# Visualisierungen in der Korpuslinguistik: diagrammatische Operationen zur Gegenstandkonstitution, -analyse und Ergebnsipräsentation

Bubenhofer, Noah

Abstract: Visualisierungen sind auch in der Korpuslinguistik wichtig, um Strukturen in großen Korpora überhaupt analysierbar zu machen. Daher sind Methoden der „Visual Analytics“ nicht einfach der letzte Schritt einer Korpusanalyse, sondern beeinflussen bereits die Datenaufbereitung. Aus Sicht der Diagrammatik, der Lehre des Diagramms, lässt sich gut herleiten, warum Visualisierungen eigentliche „Denkzeuge“ sind: Mit Diagrammen kann operiert werden und im besten Fall können aus bestehendem Wissen neue Erkenntnisse gewonnen werden. Für den korpuslinguistischen Zugang sind einige sog. diagrammatische Grundfiguren, also grundlegende Typen von Diagrammen, entscheidende Mittel, um den Untersuchungsgegenstand Sprache zu konstituieren, so z. B. die Liste, der Vektor und der Graph. Der Beitrag konzipiert diagrammatische Operationen als Grundbedingung der Korpuslinguistik und skizziert fünf diagrammatische Grundfiguren. Zusätzlich wird an einem Beispiel, der Analyse von sog. Geokollokationen, gezeigt, worin der Wert explorativer visueller Korpusanalysen besteht.

DOI: https://doi.org/10.1515/9783110538649-003

Noah Bubenhofer

# 2 Visualisierungen in der Korpuslinguistik

## Diagrammatische Operationen zur Gegenstandkonstitution, -analyse und Ergebnispräsentation

**Abstract:** Visualisierungen sind auch in der Korpuslinguistik wichtig, um Strukturen in großen Korpora überhaupt analysierbar zu machen. Daher sind Methoden der „Visual Analytics" nicht einfach der letzte Schritt einer Korpusanalyse, sondern beeinflussen bereits die Datenaufbereitung. Aus Sicht der Diagrammatik, der Lehre des Diagramms, lässt sich gut herleiten, warum Visualisierungen eigentliche „Denkzeuge" sind: Mit Diagrammen kann operiert werden und im besten Fall können aus bestehendem Wissen neue Erkenntnisse gewonnen werden. Für den korpuslinguistischen Zugang sind einige sog. diagrammatische Grundfiguren, also grundlegende Typen von Diagrammen, entscheidende Mittel, um den Untersuchungsgegenstand Sprache zu konstituieren, so z. B. die Liste, der Vektor und der Graph. Der Beitrag konzipiert diagrammatische Operationen als Grundbedingung der Korpuslinguistik und skizziert fünf diagrammatische Grundfiguren. Zusätzlich wird an einem Beispiel, der Analyse von sog. Geokollokationen, gezeigt, worin der Wert explorativer visueller Korpusanalysen besteht.

**Keywords:** datengeleitete Korpuslinguistik, Datenvisualisierung, Diagrammatik, Visual Analytics

## 1 Visualisierung als Diagramm

Die klassische Keyword in Context-Liste (KWiC) oder Konkordanz ist in der modernen Korpuslinguistik häufig nicht mal mehr der Startpunkt einer Analyse, sondern wird oft gleich in eine Tabelle oder ein Diagramm übersetzt, das die

**Anmerkung:** Der Text entstand im Rahmen des vom Schweizer Nationalfonds geförderten Projektes „Visual Linguistics", bei dem auch die folgenden Personen mitarbeiteten: Klaus Rothenhäusler, Irene Ma, Danica Pajovic und Katrin Affolter.

**Noah Bubenhofer,** Institut für Computerlinguistik, Universität Zürich, Andreasstr. 15, CH-8050 Zürich, Schweiz, E-Mail: bubenhofer@cl.uzh.ch


https://doi.org/10.1515/9783110538649-003

Verteilung eines Phänomens beschreibt. Trotzdem ist die KWiC-Liste nach wie vor ein Ikon für den korpuslinguistischen Zugang – und ich möchte im Folgenden argumentieren, dass sie weit mehr ist als das. Es handelt sich bei der Kompilation einer KWiC-Liste um eine der grundlegendsten diagrammatischen Operationen, die den korpuslinguistischen Zugang zu Sprachdaten von anderen Zugängen der Lektüre oder Analyse unterscheidet. Die KWiC-Liste ist eine Form eines Indexes oder Registers, mit dem die Einheit des Textes aufgebrochen wird. Dies ist einerseits ein Verlust, da die ursprüngliche Komplexität der Textgestalt nicht mehr (oder nur ansatzweise) sichtbar ist. Andererseits ist dies jedoch ein enormer Gewinn, weil damit eine neue Sicht auf Texte und auf die Serialität von bestimmten Phänomenen in den Texten gewonnen wird.

Es ist ein Gemeinplatz in der Korpuslinguistik, dass diese neue Perspektive auf Textkorpora gewinnbringend ist, um die Musterhaftigkeit von sprachlichen Phänomenen zu entdecken (Perkuhn & Belica 2006; Perkuhn, Keibel & Kupietz 2012). Da ich im vorliegenden Beitrag jedoch den Zusammenhang von Visualisierungen und Korpuslinguistik beleuchte, möchte ich mit einer diagrammatischen Sicht auf korpuslinguistische Analyseformen beginnen.

## 1.1 Die diagrammatische Operation als Grundbedingung der Korpuslinguistik

Die Theoriebildung zur Diagrammatik (Bauer & Ernst 2010; Krämer 2016; Stetter 2005; Bredekamp 2008) geht insbesondere auf Arbeiten von Charles Sanders Peirce zurück, der mit der Trias von Index, Ikon und Symbol das Erscheinungsspektrum von Zeichen beschrieben hat. Das prototypische Diagramm, beispielsweise die geometrische Zeichnung eines Dreiecks oder Datenpunkte in einem Koordinatensystem, steht zwar in einem ikonischen Ähnlichkeitsverhältnis zum Denotat, abstrahiert davon jedoch, so dass es als „Verwirklichung eines abstrakten Modells“ (Eco 1977: 55) wahrgenommen werden kann. Auch wenn das Dreieck nicht genau wie ein Dreieck aussieht (etwa, weil es von Hand skizziert worden ist), wird es in bestimmten Kontexten als abstraktes Modell des Dreiecks wahrgenommen. Wenn es so wahrgenommen wird, wird es als Diagramm – und nicht etwa als Skizze, Gemälde o. ä. wahrgenommen, wo die Strichdicke, Farbe, Strichqualität etc. als bedeutungstragend aufgefasst würde.[1]

**1** Natürlich können die genannten Merkmale auch bei einem Diagramm bedeutungstragend sein, etwa um verschiedene Typen von Dreiecken zu unterscheiden. Aber auch dann wird beispielsweise die Farbe Rot nicht als „rot“ wahrgenommen, sondern in ihrer diagrammatischen Funktion im Diagramm.

Mit Peirce gesprochen würden solche Merkmale nicht als „Qualizeichen", sondern als „Sinzeichen" aufgefasst (Peirce 1994: 2244).

Das Dreieck als Diagramm ist also *schematisch* und damit „auf Reproduzierbarkeit hin angelegt" (Krämer 2016: 76). Es wird in Form einer grafischen Umsetzung instantiiert und nur in dieser Instanz sichtbar.

> Alles, was schematisch ist, kann wiederholt und in dieser Wiederholung – absichtsvoll oder versehentlich – zugleich variiert werden. Dies ist ein Grundzug aller diagrammatischer Artefakte. (Krämer 2016: 77)

Zum Schematismus kommen jedoch einige weitere diagrammtypische Eigenschaften hinzu (Sybille Krämer nennt deren zwölf, Krämer 2016: 60 ff.), darunter insbesondere solche, die mit der Ausbreitung von Informationen auf einer Fläche, auf der „operiert" werden kann, zu tun haben. Diagramme benutzen interagierende *grafische Ausdrucksformen* (Punkt, Linie, Fläche, Text etc.), die auf einer *gerichteten Fläche* „unterschiedliche Gesichtspunkte und Ansichten" (Krämer 2016: 74) simultan darstellen. Diese Mittel ermöglichen es, „Relationen mit Hilfe von Relationen" (Krämer 2016: 70) darzustellen, etwa beim Dreieck, indem die Positionen der Katheten zueinander und in Relation zur Hypotenuse sichtbar werden. Denkt man an ein Punktdiagramm, wird dabei noch deutlicher, dass damit eine Verräumlichung einhergeht, in der nichträumliche Daten (Zahlenwerte) auf einer gerichteten Fläche in Relation zueinander grafisch dargestellt werden. Also: „Räumliche Relationen artikulieren – zumeist – nichträumliche Relationen" (Krämer 2016: 71), wobei Karten, Grundrisszeichnungen etc. Ausnahmen bilden. Gerade in der Korpuslinguistik haben wir es aber bei Diagrammen sehr häufig mit einer verräumlichten Darstellung zu tun – oder zumindest mit einer anders räumlich organisierten Darstellung, indem etwa die Sequenzialität von Texten aufgebrochen und vom syntaktischen in einen anders modellierten Raum überführt wird. Wenn Kollokationen beispielsweise als Netzgraph dargestellt werden, orientiert sich die Ordnung der Kollokatoren im Raum am statistischen Distributionsverhalten der Kollokationen im Korpus.

Die diagrammatisch vielleicht unscheinbare KWiC-Liste zeigt die erwähnten Eigenschaften eines Diagramms jedoch deutlich: Textfundstellen werden als Liste, damit also in Form von Zeilen ausgegeben, die in Relation zueinander, auf einer gerichteten Fläche (von oben nach unten zu lesen) stehen. Die Zeilen bilden damit einen Raum. Durch eine bestimmte Sortierung (nach Fundstellen, alphabetisch nach Kontext o. ä.) kann die Relationierung der Zeilen zueinander verändert werden. Die instantiierte KWiC-Liste verweist auf ein abstraktes, immaterielles Schema „Liste", mit dem, bei entsprechendem Wissen oder entsprechender Anleitung, die Liste interpretiert wird. Und in ihrer Form, der verräumlichten Darstellung, stellt die Liste eine Synopse dar, mit

der die an völlig unterschiedlichen Stellen auftretenden Belege simultan untereinander erscheinen.[2]

Entscheidend für die linguistische Interpretation der so dargestellten Daten ist dabei, dass mit dieser Liste – mit Sybille Krämer gesprochen – „operiert" werden kann:

> Gleich einer Karte, welche Bewegungen in einem unvertrauten Terrain eröffnet, ermöglichen Diagramme, dass wir praktisch oder theoretisch etwas tun, was ohne Diagramm schwer oder überhaupt nicht auszuführen ist. Diagramme sind graphische Denkzeuge; sie eröffnen kognitive Bewegungsmöglichkeiten, insofern ihrem Gebrauch ein transfigurativ es Potenzial innewohnt, kraft dessen graphische Konstellationen und deren handgreifliche Manipulation als intellektuelle Tätigkeiten interpretierbar werden. (Krämer 2016: 83)

Die Transfiguration der KWiC-Liste, also die räumliche, gerichtete Anordnung von Textbelegen, ermöglicht beispielsweise (bei geeigneter Datenlage) den interpretativen Schritt, daraus den Wortgebrauch und damit die Semantik eines Lexems, eine grammatische Regularität etc. abzuleiten. Das Diagramm alleine ist dafür nicht ausreichend, es benötigt auch einen theoretischen Standpunkt, von dem aus sich diese Interpretation motivieren lässt, es benötigt eine Möglichkeit, die Liste zu erstellen, aber das Schema der Liste als Diagramm ist die *conditio sine qua non*, um überhaupt eine korpuslinguistische Analyse zu ermöglichen.

Um die interpretativen Schritte zu ermöglichen, muss mit der KWiC-Liste „operiert" werden. Dies ist möglich, indem einerseits dank ihrer Diagrammhaftigkeit, in der die Fundstellen in Relation zueinander stehen und synoptisch auf einer gerichteten Fläche dargestellt werden, diese Relationen gelesen, gedeutet und manipuliert werden können. Es ist möglich, die Sortierung nach zahlreichen Kriterien zu ändern, Gruppen zu bilden und dergleichen. Andererseits stellt das Diagramm etwas dar, weist also eine Referenzialität zu etwas Gemeintem, einem diagrammexternen Sachverhalt her, überführt aber die Relationen dieses Sachverhalts in das System des Diagramms, das im Fall der KWiC-Liste dem Schema der Index-Liste gehorcht. Die Annahme, dass eine Beobachtung im Diagramm eine Parallelität zum Sachverhalt aufweist, macht die Arbeit mit dem Diagramm ja überhaupt erst sinnvoll. Man würde eine Landkarte oder einen Stadtplan nicht benutzen wollen, wenn man nicht der Überzeugung wäre, dass die Karte bzw. der Plan die wirkliche Landschaft oder Stadt

**2** Die Homophonie von „KWiC" zum englischen „quick" verleitete den Namensgeber des Terminus, Luhn (1960) (Manning & Schütze 2002: 35), wahrscheinlich auch zu diesem Akronym und verweist genau auf die Funktion der Synopsis, mit der ein schneller Überblick versprochen wird.

in einer Art und Weise abbildet, so dass man sich dank des Plans darin orientieren kann. Dies macht aber deutlich, dass Diagramme „graphische Denkzeuge“ (Krämer 2016: 83) sind – nicht nur im Fall der Korpuslinguistik an entscheidender Stelle einer ganzen Methodologie.

Die KWiC-Liste ist zwar ein bedeutendes Element korpuslinguistischer Analyse, jedoch nicht das einzige. Im Gebrauch ist eine Vielzahl von weiteren Diagrammtypen, die auf wichtige *diagrammatische Grundfiguren*, wie ich sie nennen möchte, zurückgehen. Visualisierungen in der Korpuslinguistik zu thematisieren, bedeutet vor dem Hintergrund der Diagrammatik also anzuerkennen, dass diese nicht einfach schönes Ornat für wissenschaftliche Publikationen sind, auch mehr als Analysewerkzeuge, wenn man die Perspektive der Visual Analytics hinzuzieht, sondern in ihren diagrammatischen Grundfiguren entscheidende Schritte der Gegenstands- und Methodenkonstitution, ohne die die Korpuslinguistik nicht wäre, was sie ist.

Im Folgenden möchte ich deshalb kurz auf typische Formen der Visualisierung in der Korpuslinguistik eingehen und vorschlagen, auf welche diagrammatischen Grundfiguren sie zurückzuführen sind. Dann ist es möglich herauszuarbeiten, wie diese Grundfiguren die Gegenstände und Analysemethoden konstituieren, was im vorliegenden Beitrag jedoch nur angedeutet werden kann.[3] Dieser Abschnitt gibt mir auch die Gelegenheit, auf einige wichtige Arbeiten am Institut für Deutsche Sprache (IDS) im Bereich der Datenvisualisierung aufmerksam zu machen.

Anschließend werde ich zwei Beispiele explorativer Visualisierungen zeigen. Beenden möchte ich den Beitrag mit einem Plädoyer für mehr diagrammatische Experimentierfreude.

## 1.2 Figuren der Visualisierung in der Korpuslinguistik

In der Linguistik generell, insbesondere aber auch in der Korpuslinguistik, sind m. E. folgende diagrammatischen Grundfiguren besonders relevant: Die Liste (mit allen Sonderformen wie z. B. der Tabelle), die Karte, die Vektoren, der Graph und die Partitur (vgl. Abb. 2.1). Die Typen Karte, Vektoren (z. B. in Form von Achsendiagrammen) und Graph werden in vielen Diagrammklassifizierungen genannt (so etwa im Wikipediaeintrag zu „Diagramm“), die Liste und die Partitur werden normalerweise jedoch nicht dazu gezählt. Zudem geht es mir weniger um eine formale Unterscheidung der Typen, sondern um die Frage,

**3** Ich verweise stattdessen auf Bubenhofer (2018c); Bubenhofer et al. (2017); Bubenhofer & Rothenhäusler (2016).

**Abb. 2.1:** Fünf für die Linguistik bedeutende diagrammatische Grundfiguren: Liste, Karte, Partitur, Vektoren, Graph.

wie diese Diagrammtypen sozusagen als Denkfigur in der Linguistik gegenstandskonstituierend wirken.

Im Abschnitt 1.1 habe ich die Bedeutung der KWiC-*Liste* in der Korpuslinguistik bereits betont. Generell ist die Liste eine diagrammatische Grundfigur, die für viele Wissensbereiche von großer Bedeutung ist (Echterhölter 2015; Jullien 2004; Pigeot 2004; Eco 2009). Als Index-Liste, die Fundstellen aus verschiedenen Textstellen vereint und auf die entsprechenden Quellen verweist, ist sie die Grundfigur korpuslinguistischen Arbeitens. Der Theologe und Linguist Roberto Busa war einer der Ersten, der die rechnergestützte Indexerstellung mittels IBM Lochkartenrechnern verwendete, um seinen *Index Thomesticus* zu erstellen (Busa 1951; Bonfanti 2012). Die Grundfigur erwies sich aber lange vor den ersten Computern als hilfreich, etwa für die Erstellung von Enzyklopädien (Placcius 1689; Siegel 2009).

Listen können sehr unterschiedliche Formen annehmen. Ein interessantes Beispiel ist das Kollokationsprofil, eine Liste von Lexemen, die statistisch signifikante Kollokatoren zu einem Basislexem sind. Besonders erwähnenswert ist die von Cyril Belica entwickelte Kookkurrenzberechnung (vgl. Abb. 2.2), die zwar nach dem Vorbild der Kollokationsanalyse vorgeht, jedoch weitere vielfältige Informationen in der Liste der Kollokatoren vereint: weitere sekundäre und tertiäre Kollokatoren, typische syntagmatische Muster mit Angaben zu deren Verbreitung etc. (Belica 2001 ff.).

*Karten* genießen in wissenschaftlichen Visualisierungen generell einen hohen Stellenwert und in der Linguistik sind sie z. B. in der Dialektologie bereits lange gebräuchlich, wo sie sowohl „Dokumentations-“ als auch „Forschungsmittel“ (Naumann 1982) sind. Sie dienen also sowohl der Präsentation als auch der Exploration der Daten (siehe dazu Abschnitt 1.3). Mit einer Korpusgrundlage lassen sich Karten, sofern Georeferenzen vorhanden sind, problemlos automatisch erstellen (siehe dazu auch Abschnitt 2.1). Damit werden sie für diatopische Fragestellungen zu einem wichtigen Mittel der Datenexploration, wie z. B. das Projekt „Gesprochenes Deutsch“ am Institut für

Analysewort: Ausländer, Analysetyp 0

| | | | | | | |
|---|---|---|---|---|---|---|
| -1 | -1 | 19509 | **lebenden** hier legal | 9 | 44% | von legal hier lebenden Ausländern |
| -1 | -1 | 19509 | lebenden hier rechtmäßig | 11 | 54% | sollten alle rechtmäßig hier lebenden Ausländer auch arbeiten |
| -1 | -1 | 19509 | lebenden hier | 545 | 61% | der\|die hier [...] lebenden [...] Ausländer |
| -1 | -1 | 19509 | lebenden legal | 51 | 52% | von legal [in Österreich] lebenden Ausländern |
| -1 | -1 | 19509 | lebenden rechtmäßig | 33 | 51% | von rechtmäßig [in Deutschland] lebenden Ausländern allein abgeschoben |
| -1 | -1 | 19509 | lebenden | 1887 | 66% | in hier\|Deutschland lebenden [...] Ausländer |
| -2 | -2 | 13649 | **Integration** lebender | 34 | 100% | die Integration hier\|in Österreich lebender Ausländer |
| -2 | -2 | 13649 | Integration Aussiedlern | 24 | 75% | die\|zur Integration von Ausländern und Aussiedlern |
| -2 | -2 | 13649 | Integration | 1933 | 62% | die Integration von Ausländern |
| -2 | -2 | 13225 | **Ausländerinnen** erleichterte | 17 | 52% | erleichterte Einbürgerung junger Ausländerinnen und Ausländer |
| -2 | -2 | 13225 | Ausländerinnen Schweiz | 45 | 60% | Ausländerinnen und Ausländer in\|die der Schweiz |
| -2 | -2 | 13225 | Ausländerinnen Stimm | 15 | 53% | Ausländerinnen und Ausländern ... das Stimm und Wahlrecht |
| -2 | -2 | 13225 | Ausländerinnen | 950 | 67% | Ausländerinnen [und] Ausländer |
| -4 | 5 | 7413 | **Deutschland** lebende | 228 | 98% | in Deutschland [...] lebende [...] Ausländer |
| -4 | 5 | 7413 | Deutschland geborene | 123 | 57% | in\|In Deutschland geborene Kinder von Ausländern die |
| -4 | 5 | 7413 | Deutschland geborenen | 86 | 82% | in Deutschland [...] geborenen [Kinder\|Kindern von] Ausländern die ... |
| -4 | 5 | 7413 | Deutschland | 3688 | 42% | Ausländer [... in] Deutschland |
| -5 | 4 | 6904 | **illegal** eingereiste | 130 | 96% | illegal [...] eingereiste [...] Ausländer |
| -5 | 4 | 6904 | illegal eingereisten | 41 | 63% | von illegal [...] eingereisten [...] Ausländern |
| -5 | 4 | 6904 | illegal beschäftigte | 72 | 90% | illegal [...] beschäftigte Ausländer |
| -5 | 4 | 6904 | illegal | 1107 | 58% | illegal [in ...] Ausländer |
| -1 | -1 | 6673 | **lebende** hier legal | 5 | 100% | legal hier lebende Ausländer |
| -1 | -1 | 6673 | lebende hier | 153 | 96% | für hier [...] lebende [...] Ausländer |
| -1 | -1 | 6673 | lebende legal | 36 | 100% | legal [in Österreich] lebende Ausländer |
| -1 | -1 | 6673 | lebende | 744 | 97% | in hier\|Deutschland lebende [...] Ausländer |
| -1 | -1 | 6022 | **viele** leben | 143 | 62% | zu viele [...] Ausländer [...] leben |
| -1 | -1 | 6022 | viele lebten | 29 | 62% | es\|Deutschland lebten zu viele [...] Ausländer in |
| -1 | -1 | 6022 | viele wohnen | 49 | 67% | in dem viele [...] Ausländer [...] wohnen |
| -1 | -1 | 6022 | viele | 1867 | 84% | viele [...] Ausländer |

**Abb. 2.2:** Kookkurrenzprofil von „Ausländer“ (Ausschnitt), Kookkurrenzdatenbank CCDB (Belica 2001 ff.).

Deutsche Sprache zeigt (vgl. Abb. 2.3; Kleiner 2011 ff.). Die Kartendarstellung als diagrammatische Grundfigur fügt zu sprachlichen Äußerungen eine weitere Dimension, nämlich eine geografische, hinzu.

Die diagrammatische Grundfigur der *Partitur* entstand in Ansätzen bereits im neunten Jahrhundert, um die Polyphonie von Musik sichtbar zu machen (Sachs & Röder 1989), also die Gleichzeitigkeit von Stimmen. Offensichtlich hat diese Idee der Notation in den 1970er Jahren Eingang in die Gesprächsanalyse gefunden (Sacks, Schegloff & Jefferson 1974; Redder 2001), wo sie ein wichtiges Element war, um gesprochene Sprache überhaupt unter einer neuen Perspektive, die den Turn in den Fokus nimmt, zu konstituieren (vgl. für weitere Ausführungen dazu Bubenhofer 2018c). In der Korpuslinguistik ist diese Grundidee jedoch viel unscheinbarer auch in Annotationssystemen enthalten: Eine Auszeichnungssprache wie XML fügt genauso Ebenen oder „Stimmen“ zu einem

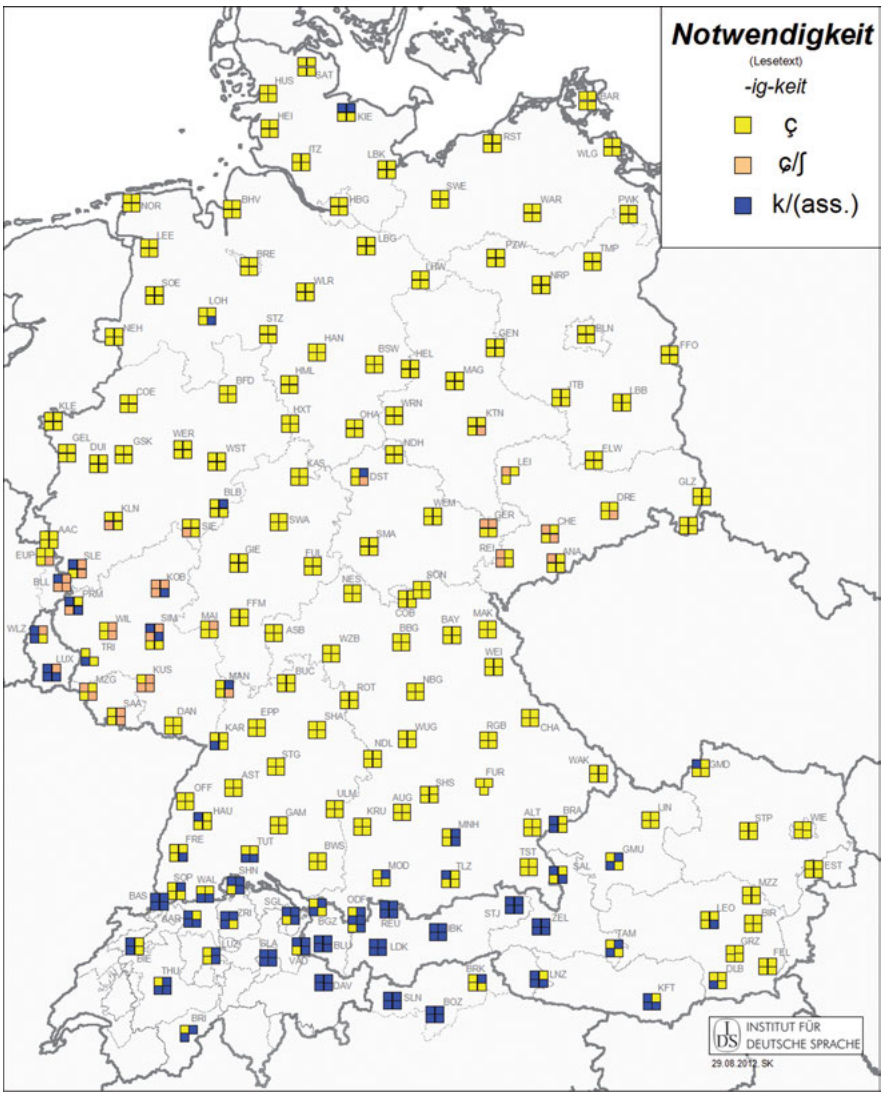


**Abb. 2.3:** Realisierung des auslautenden Konsonanten <ig> im abgeleiteten Abdjektivabstraktum *Notwendigkeit*, http://prowiki.ids-mannheim.de/bin/view/AADG/IgKeit (letzter Zugriff: 6. 11. 2017), (Kleiner 2011 ff.).

Primärtext hinzu, wie das die Partitur macht.[4] Auch die Darstellung von Korpora als „vertikalisierter" Text, indem pro Zeile ein Token und durch Tabulator oder ein anderes Trennzeichen separiert weitere Informationen zum Token (Wortart, Grundform etc.) als „Spalten" hinzugefügt werden, nimmt auf die Grundfigur der Partitur Bezug. Die diagrammatische Grundfigur der Partitur erlaubt es also, Text zwar als sequenziell, jedoch als komplexes Mehr-Ebenen-Phänomen darzustellen und fügt zu dem Text weitere Dimensionen hinzu.

Die Grundfigur der *Vektoren* ist in der Korpuslinguistik besonders relevant: Um die gängigen Formen der Darstellung von Linien-, Balken-, Punktdiagrammen etc. zu ermöglichen, ist vorher die Transformation von Sprache in einen Vektorraum notwendig. Ausgehend von einer Index-Liste von Belegen, werden diese in eine Tabelle von Vektoren überführt, wo beispielsweise Frequenzen nach Jahren geordnet werden, um ein Liniendiagramm zur diachronen Entwicklung eines Phänomens zu erstellen (vgl. dazu etwa die Arbeiten des Lexik-Projekts „Empirische Methoden" am IDS, z. B. zu diachronem Korpuswandel, Koplenig 2017). Eine Erweiterung ist die Idee, komplexere Vektoren zu erstellen, etwa Kollokationsprofile: Diese werden als Vektor repräsentiert, der die Frequenzen des Kovorkommens des Basislemmas mit den Kollokatoren enthält. Dadurch wird die Semantik (wenn man einem Begriff der distributionellen Semantik folgt) als Vektor codiert und damit mit anderen Vektoren verrechenbar: Es lassen sich Abstände, Homologien etc. berechnen, wie dies etwa bei Verfahren des Word Embeddings geschieht (Mikolov et al. 2013).[5] In gleicher Weise, aber auf Ebene von Texten, wird beim Topic Modelling verfahren (Graham, Weingart & Milligan 2012). Die diagrammatische Grundfigur der Transformation sprachlicher Daten in Vektoren ist der eigentliche Schritt der „Verdatung" von Sprache, um sie mathematischen Operationen zugänglich zu machen (Bubenhofer & Rothenhäusler 2016: 63) und damit am Beispiel der distributionellen Semantik eine Wortbedeutung zu modellieren, die eine „*fiktive*, mathematisch aus Beobachtungsdaten erzeugte Entität" (Bender & Marrinan 2014: 197) darstellt.

*Graphen* in gerichteter oder ungerichteter Form gehören ebenfalls zu sehr alten Formen von Diagrammen, die etwa in Form von Stammbäumen bereits vor über 2000 Jahren gezeichnet wurden (Kruja et al. 2002; Lima 2014). In der

**4** XML leistet darüber hinaus aber natürlich noch mehr, indem beispielsweise auch eine Hierarchisierung der Ebenen zueinander abbildbar ist. Vgl. aber Bański (2010) gerade für die Probleme von XML zur Annotation von Text.

**5** Vgl. für eine Kombination von Word Embeddings und diachronem Vergleich die Anwendung „DiaViz" von Peter Fankhauser und Marc Kupietz (IDS): http://corpora.ids-mannheim.de/diaviz/dereko.html (letzter Zugriff: 6. 11. 2017)

Korpuslinguistik werden Graphen insbesondere für die Visualisierung von Kollokationen verwendet (Bubenhofer 2018b; Brezina, McEnery & Wattam 2015), so auch in der CCDB des IDS (Belica 2001 ff.) oder im Rahmen des Projekts „Usuelle Wortverbindungen“ (Steyer 2013). Aus diagrammatischer Sicht zeichnet sich die Graphdarstellung dadurch aus, dass die strukturelle Information darüber, welche Knoten Verbindungen zueinander aufweisen, von der darstellerischen entkoppelt ist: Die Definition der Knoten-Kanten-Beziehungen muss für die Darstellung als Netzwerkgraph in ein Layout überführt werden, also meist ein algorithmisch definierbares Prinzip, nachdem die Knoten angeordnet werden. Im Kontext der Netzwerkanalysen werden oft „force-directed“-Layoutmodi verwendet, die die Knoten nach den physikalischen Prinzipien der Anziehung und Abstoßung in einem energetischen Optimum platzieren. Knoten, die viele Verbindungen untereinander haben, tendieren dann dazu, nahe beieinander zu stehen. Ein solches Bild ist semantisch (bei Kollokationen), hermeneutisch (z. B. bei typischen Wortverbindungen in Geschichten o. ä.) etc. interpretierbar.

## 1.3 Präsentation vs. Exploration

Eine wichtige Unterscheidung von Visualisierungstypen generell, die quer zu den oben beschriebenen Typen steht, ist jene von Präsentations- und Explorationsgrafiken (Chen, Härdle & Unwin 2008: 4–5; Schumann & Müller 1999: 5). Erstere dienen dazu, der Forscherin oder dem Forscher bereits bekannte Erkenntnisse in einem Diagramm darzustellen, etwa um sie besser lesbar zu machen. Explorative Visualisierungsmethoden hingegen werden im Analyseprozess eingesetzt, um die Daten überhaupt interpretierbar zu machen. Das Paradigma der „Visual Analytics“ (Keim et al. 2010; Chen, Härdle & Unwin 2008) nutzt solche Techniken und verbindet so quantitativ-maschinelle mit qualitativ-interpretierenden Analysemethoden:

> Visualisation becomes the medium of a semi-automated analytical process, where humans and machines cooperate using their respective, distinct capabilities for the most effective results (Keim et al. 2010: 14).

Im Anschluss an die diagrammatischen Überlegungen in Abschnitt 1.1 lässt sich sagen, dass explorative Visualisierungen Diagramme mit ausgeprägter Operationalität sind: Im besten Fall nutzen explorative Visualisierungen ein Diagrammschema, mit dem die Daten so dargestellt werden, dass dank der Darstellung neue Erkenntnisse gewonnen werden können. Dies ist beispielsweise der Fall, wenn Kollokationen als Netz dargestellt werden, wie in Ab-

schnitt 1.2 zur diagrammatischen Grundfigur Graph bereits ausgeführt worden ist (Pfeffer 2010; Kruja et al. 2002; Chen, Härdle & Unwin 2008: 109; Bubenhofer 2018b).

Die in einem Netzwerkgraph sichtbar gewordenen Cluster von Knoten sind nun eine Erkenntnis, die aufgrund der Listen alleine wohl nicht hätte gewonnen werden können (vor allem bei großen Datenmengen). Das Beispiel macht jedoch auch deutlich, dass nicht nur die grafische Umsetzung alleine den Mehrwert für die Analyse ermöglichte, sondern Darstellungsprinzipien generellerer Natur: das Netz, die physikalischen Prinzipien, die statistische Berechnung etc.

Mit einem weiten Diagramm-Begriff könnte man argumentieren, dass diese generellen Darstellungs- und Ordnungsprinzipien (wie Listen, Matrizen, Vektorräume etc.) zum grafischen Diagramm dazu gehören und Operationalität im umfassenden Sinn ermöglichen. Es ist ein Spezifikum von Digitalität, dass die verschiedenen Formen der Materialisierung von Daten (als Bildschirmbild, Ausdruck etc.) mit den immateriellen Repräsentationen und Manipulationen von Daten einhergehen und darauf aufbauen (Bubenhofer 2018c). Diagrammatische Operationen umfassen deshalb mehr als Interaktivität in einem Diagramm; dazu gehören auch die Operationen der immateriellen Repräsentation und Manipulation von Daten.

Bei einem engen Diagramm-Begriff würde man die digitalen Operationen nicht als diagrammhaft verstehen, sondern erst die grafische Repräsentation des Diagramms. Doch auch dann ist augenfällig, dass insbesondere beim digitalen Diagramm die Verbindung zu den Daten so eng ist, dass eine scharfe Trennung zwischen Diagramm und Nicht-Diagramm schwierig ist.

Präsentationsgrafiken sind oft weniger offensichtlich operational. Ein Rest an Operationalität ist aber auch da vorhanden, da sie genauso als „Denkzeuge“ (Krämer 2016: 83) fungieren. Die Darstellung eines Ausschnitts aus einer KWiC-Liste in einer Publikation legt eine bestimmte (von der Erstellerin erwünschte) Interpretation nahe, verhindert aber nicht, dass der Rezipient damit noch andere Erkenntnisse gewinnt (auch wenn sie unvollständig sein mögen), die der Erstellerin entgangen sind oder nicht im Fokus standen.

## 2 Explorative Visualisierungen: Anwendungsbeispiele

Die bisherigen Ausführungen waren theoretischer Natur. Im Folgenden möchte ich in aller Kürze von zwei Experimenten berichten, bei denen wir visuelle

Analysemethoden einsetzten, um Korpusdaten besser zu verstehen. Im vorliegenden Beitrag geht es mir um die diagrammatischen Komponenten und die damit zusammenhängenden Reflexionen.

## 2.1 Geokollokationen

Die Berechnung und Darstellung von „Geokollokationen" zielt darauf ab, diskursiv geprägte Konstruktionen von Welt zu beschreiben (Bubenhofer 2014; Bubenhofer et al. 2017). In Diskursen entstehen bestimmte Assoziationen zu Orten und Regionen: *Schweiz – Banken*, *Schokolade*, *Steuerhinterziehung*; *Griechenland – Finanzkrise*, *Urlaub*, *Flüchtlingskrise* etc. Das Ziel des Experiments ist es, datengeleitet aus Korpora solche Assoziationen abzuleiten.

Als Datenbasis dienten uns verschiedene Quellen (Presse, Bundestagdebatten, Webdiskussionsforen etc.). Die im Folgenden präsentierten Daten beruhen auf einem Korpus von Artikeln des Magazins *Der Spiegel* und der Wochenzeitung *Die Zeit* von 1946 bis 2016 (611 Mio. Tokens). Die Daten wurden mit dem TreeTagger (Schmid 1994, 1995) und der Standardbibliothek für Deutsch lemmatisiert und mit Wortarten nach dem Stuttgart-Tübingen-Tagset (Schiller, Teufel & Thielen 1995) annotiert. Zusätzlich wurde der Stanford Named Entity Recognizer (NER) (Finkel, Grenager & Manning 2005) in einer fürs Deutsche adaptierten Version (Faruqui & Padó 2010) auf die Daten angewandt.

Ausgehend von den vom NER-Tagger als Toponyme erkannten Lexeme wurden die zu dieser Basis statistisch auffälligen Kollokatoren im gleichen Satz mittels Log Likelihood Signifikanztest (Evert 2009) berechnet. Die daraus resultierenden Listen sind lang und umfassen Toponyme und damit kollokierende Lexeme. Abbildung 2.4 zeigt einen Ausschnitt aus einer solchen Liste, die jedoch bereits mit weiteren Informationen, darunter Georeferenzen angereichert ist.

Um die Unübersichtlichkeit der Liste zu beheben, ist es naheliegend, die Daten auf eine Karte zu projizieren. Dafür ist die Georeferenzierung – also die Anreicherung der Toponyme mit Geokoordinaten – notwendig, der wiederum oft eine Disambiguierung vorausgehen muss. Für die Disambiguierung verwenden wir den „Cartographic Location and Vicinity Indexer" (CLAVIN),[6] der Ortskandidaten nach Populationsgröße ordnet (*Berlin* in Deutschland vor *Berlin* in den USA) und danach den Textkontext berücksichtigt und Orte bevorzugt, die nahe beieinander liegen.

**6** Vgl. https://github.com/Berico-Technologies/CLAVIN/ (letzter Zugriff: 6. 11. 2017).

```
name lon lat type country cow_code freq sig word pos
"Yunnan" 102 25 state cn 31 0.0001 "Provinz" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 27 0.0001 "töten" VVPP
"Sinai" -78.1885830257228 -2.08513105 administrative ec 32 0.0001 "werden" VAPP
"Sinai" -78.1885830257228 -2.08513105 administrative ec 36 0.0001 "Al-Arisch" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 23 0.0001 "Islamisten" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 54 0.0001 "Halbinsel" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 24 0.0001 "Mann" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 30 0.0001 "öffentlich" ADJA
"Sinai" -78.1885830257228 -2.08513105 administrative ec 21 0.0001 "Al-Arisch" ADJD
"Sinai" -78.1885830257228 -2.08513105 administrative ec 20 0.0001 "Demonstrant" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 39 0.0001 "Norden" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 24 0.0001 "Polizist" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 25 0.0001 "Stadt" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 29 0.0001 "ägyptisch" ADJA
"Sinai" -78.1885830257228 -2.08513105 administrative ec 26 0.0001 "Extremist" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 30 0.0001 "Gebäude" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 36 0.0001 "bewaffnet" ADJA
"Dominikanische Republik" -70.3012705 19.094175 administrative do 68 0.0001 "Republik" NN
"Dominikanische Republik" -70.3012705 19.094175 administrative do 50 0.0001 "Dominikanische" NN
"Katar" 51.2295295 25.3336984 administrative qa 63 0.0001 "Präsident" NN
"Katar" 51.2295295 25.3336984 administrative qa 56 0.0001 "Todesfall" NN
"Katar" 51.2295295 25.3336984 administrative qa 31 0.0001 "Fußball-Bund" NN
"Katar" 51.2295295 25.3336984 administrative qa 49 0.0001 "Menschenrecht" NN
"Katar" 51.2295295 25.3336984 administrative qa 22 0.0001 "Temperatur" NN
"Katar" 51.2295295 25.3336984 administrative qa 50 0.0001 "Turnier" NN
"Katar" 51.2295295 25.3336984 administrative qa 29 0.0001 "geplant" ADJA
"Katar" 51.2295295 25.3336984 administrative qa 24 0.0001 "alarmierend" ADJA
"Katar" 51.2295295 25.3336984 administrative qa 74 0.0001 "international" ADJA
"Katar" 51.2295295 25.3336984 administrative qa 24 0.0001 "Situation" NN
```

**Abb. 2.4:** Ausschnitt aus einer Liste von Geokollokationen: In der ersten Spalte das Toponym, in den letzten beiden Spalten der Kollokator mit Wortartklasse.

Nun ist es möglich, die Daten auf einer Karte darzustellen. Dafür verwenden wir Javascript und insbesondere die Bibliothek „D3“ (Bostock, Ogievetsky & Heer 2011). Abbildung 2.5 zeigt einen Ausschnitt aus der Kartendarstellung: Grundsätzlich werden Orte, an denen Kollokatoren vorhanden sind, als Punkte dargestellt, wobei die Punktgröße die Anzahl der Kollokatoren repräsentiert. Je nach Einstellung im Kontrollfeld der Karte werden anstelle der Punkte immer (oder nur bei genügend vorhandenem Platz) die Kollokatoren direkt als Text angezeigt.

Die Geokollokationen-Visualisierung ist ein exploratives Werkzeug und umfasst deshalb auch ein Kontrollfeld mit verschiedenen Einstellmöglichkeiten, um die Daten zu filtern: Auswahl des Datensatzes (des Korpus), Setzen von Schwellwerten, Restriktion auf bestimmte Wortartklassen und eine Einschränkung auf Kollokatoren, die auf einen regulären Ausdruck passen. So zeigt Abbildung 2.6 Orte und Staaten, die im Zusammenhang mit *Flucht*, *Flüchtling* oder *Migration* genannt werden.[7]

**7** Gesucht wurde nach Kollokatoren, die auf den regulären Ausdruck `.*([Ff]l[uü]cht|[Mm]igra).*` passen.

**Abb. 2.5:** Ausschnitt aus der Kartendarstellung der Geokollokationen: Anzeige von Orten mit Kollokatoren als Punkte und mit Text bei genügend Platz.

**Abb. 2.6:** Ausschnitt aus der Kartendarstellung der Geokollokationen: Restriktion auf Kollokatoren, die die Zeichenkette *flucht*/*flücht* oder *migra* enthalten, also zu den Themen *Flüchtlinge* und *Migration* – Punktgröße repräsentiert die Anzahl der Kollokatoren.


**Abb. 2.7:** Ausschnitt aus der Kartendarstellung der Geokollokationen: Restriktion auf Kollokatoren, die die Zeichenkette *flucht*/*flücht* oder *migra* enthalten, also zu den Themen *Flüchtlinge* und *Migration* – Korpus Zeit/Spiegel 2010–2016, Anzeige der Kollokatoren.

Die Kartendarstellung erlaubt es nun in der Folge, geografische Zusammenhänge zwischen den Toponymen und ihren Assoziationen zu entdecken. So wird sichtbar, dass bestimmte Kollokatoren sehr global verwendet werden, z. B.: *Stadt*, *Land* oder *Jahr*. Andere hingegen sind spezifisch für bestimmte Regionen, die eine Gemeinsamkeit aufweisen, wie beispielsweise *Menschenrecht* oder *Flüchtling*. Andere sind sehr ortsspezifisch wie z. B. *chinesisch* oder *Obama*.

Schränkt man die Anzeige der Kollokatoren ein, um einen thematisch definierten Diskurs zu untersuchen, lassen sich die Nuancen der Berichterstattung dazu und die damit konstruierten geografischen Assoziationen entdecken. In Abbildung 2.7 sind die relevanten Kollokatoren im Bereich Flucht/Migration mit der selben Einschränkung wie in Abbildung 2.6 im Raum Deutschland, Balkan, Türkei, Naher Osten sichtbar. Auffallend sind die vielen Derivationen (hauptsächlich Nominalkomposita) von den Lexemen *Flüchtling*, *Flucht* und *Migration* wie etwa *Flüchtlingslager*, *Zuflucht*, *Bootsflüchtling*, *Flüchtlingswelle*, *Flüchtlingszahl* etc. Allerdings werden diese Derivationen hauptsächlich im

Zusammenhang mit Deutschland und Europa (gemeint ist dabei meist die Europäische Union) genannt, nicht mit den Ursprungsländern der Flucht. Dort gibt es generell weniger Derivationen. Dies deutet darauf hin, dass im Diskurs Migration primär als innenpolitisches Thema konstruiert wird – in der deutschen Presse als deutsches und EU-Problem – und nicht als Problem der Ursprungsländer oder der Länder, die außerhalb Deutschlands als Transitländer davon betroffen sind.

Dank der Korpusdaten von *Der Spiegel* und *Die Zeit*, die die ganze Nachkriegszeit abdecken, können auch diachrone Veränderungen analysiert werden. Wechselt man etwa bei der gleichen Einschränkung der Kollokatoren auf den Migrationsdiskurs die Datengrundlage und wählt anstelle des Zeitraums 2010 bis 2016 die Nachkriegsjahre, werden die Unterschiede alleine anhand der damit assoziierten Regionen sichtbar. Anstelle Afrikas wird der ganze amerikanische Kontinent damit verknüpft, was natürlich daran liegt, dass aus deutscher Perspektive Migration in der Nachkriegszeit auch ein Emigrationsthema war.[8]

Die Darstellung der Geokollokationen auf einer Karte ist naheliegend und unterstützt eine Denkfigur, die in unseren Köpfen wahrscheinlich automatisch anspringt, wenn wir Toponyme lesen: die geografische Verortung. Diese Denkfigur ist geprägt von den uns bekannten Kartenbildern. Karten sind jedoch immer zweidimensionale Projektionen der kugelartigen Welt, die immer verzerrt ist und beispielsweise bei der uns vertrauten Mercator-Projektion die Länder am Äquator im Vergleich zu den davon entfernten Ländern viel kleiner darstellt (Glasze 2009; Smith 1992). Zudem stellt sich das Problem, dass die Größe der abgebildeten Länder geografisch vorbestimmt ist und nicht zwingend auch die diskursive Bedeutung widerspiegelt. Daher ist es gerade interessant, unterschiedliche Visualisierungslösungen auszuprobieren – und damit mit dem Diagramm ein anderes System der Relationen zu konstruieren.

Dafür implementierten wir eine sog. Dorling-Ansicht (Dorling 1993), bei der die geografischen Entitäten (z. B. Staaten) als Punkte dargestellt werden, deren Größe eine Datenvariable repräsentiert. Die Positionierung der Punkte folgt zwar einer geografischen Ordnung, geht aber zwangsweise Kompromisse ein, um Überlappungen zu vermeiden. Die Größe der Punkte repräsentiert in unserem Fall die Anzahl signifikanter Kollokatoren zum jeweiligen Staat.

Diese Darstellung kombinierten wir mit einem Sankey-Diagramm (Sankey 1896), das Flüsse zwischen Entitäten in Form von unterschiedlich breiten Linien darstellt. In unserem Fall nutzten wir diese Darstellung, um eine (von der Benutzerin/dem Benutzer) ausgewählte Zahl von Kollokatoren unterhalb

**8** Siehe für weiterführende Ausführungen dazu auch Bubenhofer et al. (2017).

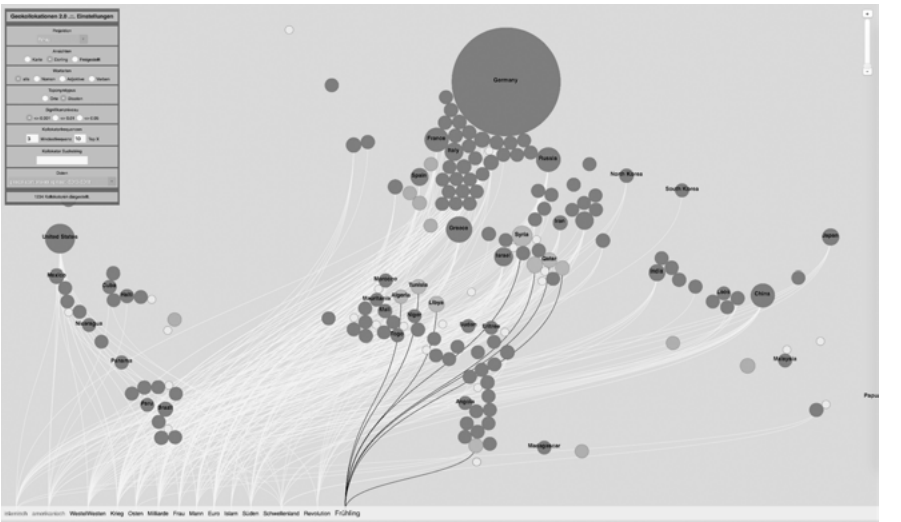


**Abb. 2.8:** Kombinierte Darstellung der Karte als Dorling- und Sankey-Diagramm. Ausgewählt ist der Kollokator *Frühling*, der auf die Länder, die mit dem *Arabischen Frühling* assoziiert werden, verweist.

der Karte darzustellen und anzuzeigen, mit welchen Ländern diese kollokieren. Abbildung 2.8 zeigt diese Ansicht.

Dieser Versuch, von der geografischen Darstellung zu abstrahieren, kommt einer Darstellung der diskursiv geprägten Weltsicht näher als die konventionelle Karte. Sie greift die Diskussionen der kritischen Kartographie (Glasze 2009) auf und deutet an, dass es immer eine Vielzahl von Möglichkeiten gibt, das ikonische Ähnlichkeitsverhältnis zwischen Denotat und Diagramm zu konstruieren.

## 2.2 Narrative

Im Fall der Geokollokationen (in Abschnitt 2.1) ist die Darstellung der Daten auf einer Karte eine naheliegende Lösung, die dabei hilft, Bezüge zwischen den Datenpunkten sichtbar zu machen. Beim zweiten Beispiel für explorative Visualisierung geht es darum, überhaupt erst adäquate Formen der Visualisierung zu finden, um die Daten analysierbar zu machen. Es geht darum, typische narrative Muster in seriellen Alltagserzählungen (Ehlich 1980; Gülich 1980; Quasthoff 1980) zu finden. So gibt es beispielsweise Webdiskussionsforen, in denen Frauen die Geburt ihrer Kinder erzählen. Für solche Erzählungen wird meist ein formaler Rahmen innerhalb allgemeiner Foren zu Elternschaft bereit-

**Tab. 2.1:** Korpus „Geburtsberichte“.

| | # Wörter | # Texte |
|---|---|---|
| http://www.urbia.de/ | 7.364.108 | 8.808 |
| http://www.babyforum.de/ | 2.089.936 | 1.824 |
| http://www.parents.at/ | 1.199.174 | 1.647 |
| https://www.swissmomforum.ch/ | 1.156.193 | 919 |
| http://www.eltern.de/ | 438.017 | 716 |
| http://www.umstandsforum.de/ | 289.807 | 568 |
| **Total** | **12.537.235** | **14.482** |

gestellt, meist ein Unterforum mit dem Titel „Geburtsberichte“ o. ä. Dort schreiben die Mütter dann in Form von Postings ihre Erlebnisse und andere Leserinnen[9] kommentieren diese Initialpostings. Auffallend ist eine deutliche Serialität der Erzählungen: Obwohl das Geburtserlebnis individuell einmalig ist, sind es Geburten überhaupt nicht. Und auch die Erzählungen darüber folgen offensichtlich bestimmten Folien, die soziokulturell verankert sind.

Ich werde an dieser Stelle nur kurz die Eckdaten der Studie skizzieren und hauptsächlich die diagrammatischen Aspekte problematisieren. Im Detail wird die Studie in Bubenhofer (2018a) vorgestellt.[10]

Ziel der Studie ist es, solche „narrativen Muster“ der Geburtserzählung (in den erwähnten Foren) datengeleitet zu berechnen. Dafür stellte ich ein Korpus von über 14.000 Geburtsberichten (12 Mio. Tokens) aus sechs Foren im deutschsprachigen Raum zusammen (vgl. Tab. 2.1).

Die Korpusdaten wurden mit Hilfe des TreeTaggers (Schmid 1994) lemmatisiert und mit dem Stuttgart-Tübingen-Tagset (Schiller, Teufel & Thielen 1995) mit Wortartklassen annotiert. Um typische Formulierungsmuster in den Berichten zu finden, verwendeten wir den bereits an verschiedenen Daten erprobten Ansatz, typische n-Gramme zu extrahieren (Bubenhofer 2017). Dazu benutzten wir ein Referenzkorpus von Presseartikeln aus *Die Zeit* und *Der Spiegel* – das gleiche Korpus wie in Abschnitt 2.1 angegeben, allerdings nur den Zeitraum von 2010 bis 2016 – und berechneten in jedem Korpus alle auftretenden Wort-n-Gramme mit $n > 5$. Für die Berechnung berücksichtigten wir jedoch jeweils die Grundformen (Lemmata) anstelle der Wortformen. Zudem durfte sich ein n-Gramm nicht über einen Satz hinaus erstrecken. Anschließend verglichen

**9** Es scheint sich fast immer um Frauen zu handeln.

**10** Vgl. für eine breite Einführung in die Erzählforschung zu Geburtsberichten Colloseus (2016) und für eine kleinere linguistische Studie dazu Barbieri et al. (2012).

wir die Frequenzen der n-Gramme in den beiden Korpora und führten einen Log Likelihood-Signifikanztest durch, um die n-Gramme, die typisch für die Geburtsberichte sind, zu extrahieren (inkl. jener, die nur im Geburtsberichte-Korpus vorkommen).

Soweit verfolgt die Methode einen sog. „Bag of Words"-Ansatz, bei dem zwar nicht Einzellexeme, sondern n-Gramme, jedoch ohne Rücksicht auf ihre

Abfolgen und Positionen in den Geschichten, erfasst werden. Um narrative Muster aufdecken zu können, ist es jedoch notwendig, typische Sequenzen von n-Grammen zu finden. Daher nutzten wir folgende zwei Strategien:

- Zu jedem n-Gramm-Token wurde die relative Position in der Geschichte (zwischen 0 – Anfang – und 1 – Ende der Geschichte) erfasst, so dass pro n-Gramm-Type der Mittelwert und die Standardabweichung der Positionen berechnet werden kann.
- Weiter berechneten wir für jeden n-Gramm-Type die damit links und rechts kollokierenden n-Gramm-Types, in Anlehnung an den in Bubenhofer, Müller & Scharloth (2013) beschriebenen Ansatz.[11]

Abbildung 2.9 zeigt im Überblick die für die Geburtsberichte signifikanten n-Gramme und ihre relativen Positionen in den Geschichten in Korrelation zur Standardabweichung der Positionen. Dabei ist eine deutliche Korrelation zwischen Position und Variation ersichtlich: Die n-Gramme am Anfang und am Ende der Geschichten sind in ihren Positionen relativ stabil, während in der Mitte der Geschichten mehr Variation in der Position zu beobachten ist. Das sehr häufige n-Gramm *das Licht der Welt erblickt* kommt zwar im Mittel tatsächlich auch etwa in der Mitte der Geschichte vor, die Standardabweichung ist jedoch relativ hoch, da es oft auch eher am Anfang oder Ende der Geschichten auftritt. Ein n-Gramm wie *der eigentliche ET war der*[12] steht jedoch in fast allen Geschichten jeweils am Anfang.

Es gibt nun verschiedene Kriterien, nach denen die n-Gramme angeordnet werden können. Die relative Position in der Geschichte ist dabei das wichtigste Kriterium. Frequenz, Standardabweichung der Position aber auch Ähnlichkeit der n-Gramme zueinander sind aber weitere wichtige Kriterien. Um eine flexible Exploration der Daten zu ermöglichen, experimentierten wir mit interaktiven, dreidimensionalen Darstellungen (vgl. das Bildschirmfoto in Abb. 2.10). Dreidimensionale Diagramme bedürfen der Interaktion und können

**11** Die genaue Implementierung ist in Affolter (2016) beschrieben.

**12** „ET" steht für „errechneter Termin".

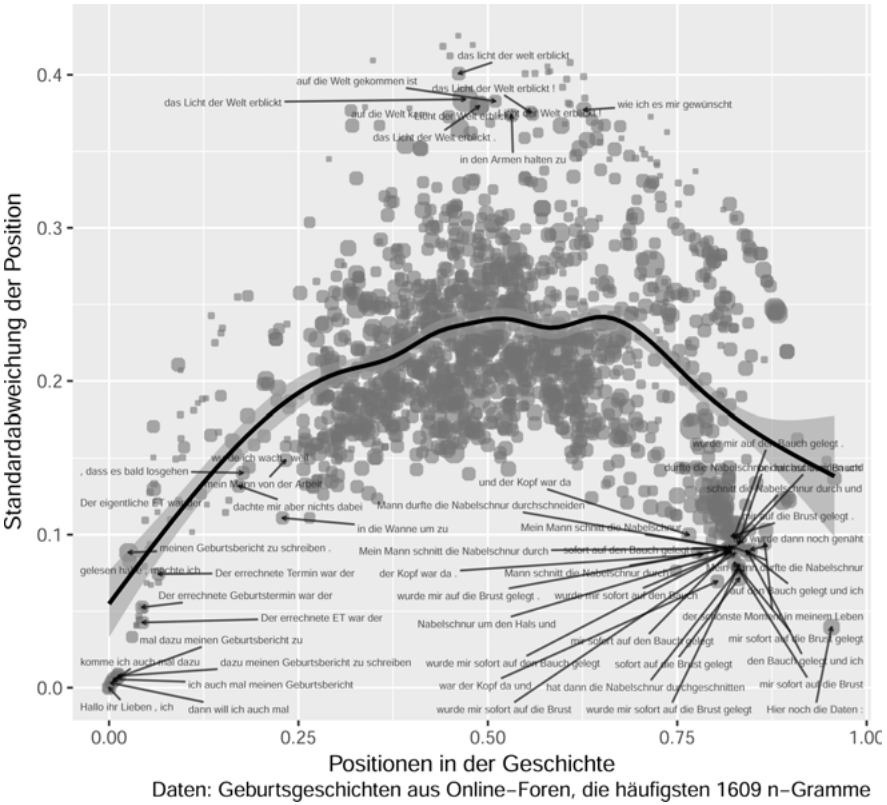


**Abb. 2.9:** Positionen der vorhandenen n-Gramme in den Geburtsberichten in Korrelation zur Standardabweichung der Position (Abb. aus Bubenhofer 2018a).

auf Papier nur unzureichend dargestellt werden, weswegen ich auf die Online-Version verweise.[13]

Der Vorteil von interaktiven, dreidimensionalen Darstellungen ist, dass je nach Bedarf verschiedene Perspektiven auf die Daten möglich sind. Bei der Vorderansicht (x-Achse von links nach rechts, y-Achse von unten nach oben) wie in Abbildung 2.10 wird die Korrelation von Position und Frequenz fokussiert, eine Draufsicht (z statt y-Achse) zeigt die verschiedenen Varianten ähnlicher n-Gramme. Mit den zahlreichen „Zwischenperspektiven" können die Ordnungskriterien beliebig gewichtet werden.

**13** Vgl. https://www.bubenhofer.com/sprechtakel/2017/02/19/die-serielle-singularitaet-vierzehntausend-geburtsgeschichten/ (letzter Zugriff: 6. 11. 2017).

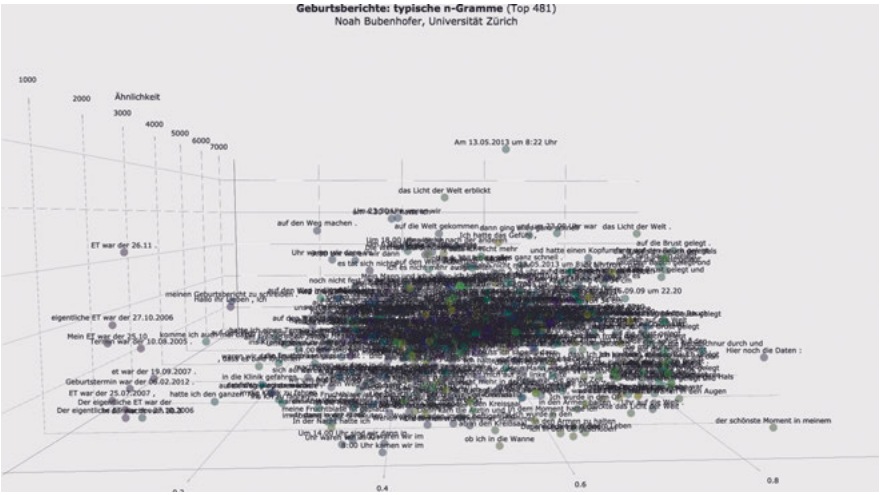


**Abb. 2.10:** Dreidimensionale Darstellung der Gebuchtsberichte-n-Gramme – x-Achse: Relative Position; y-Achse: Frequenz; z-Achse: Ähnlichkeit (Abb. aus Bubenhofer 2018a).

In den Darstellungen in Abbildung 2.9 und Abbildung 2.10 sind die Kollokationsinformationen, also die Angaben darüber, welche n-Gramm-Sequenzen typischerweise vorkommen, nicht dargestellt. Um diese Informationen darstellbar zu machen, entwickelten wir das Visual-Analytics-Tool „NarrViz" (implementiert von Katrin Affolter, vgl. Affolter 2016). Abbildung 2.11 zeigt die Oberfläche mit dem geladenen Datensatz der Geburtsberichte. Es handelt sich um eine in Javascript realisierte Webbrowser-Anwendung.[14]

Auch in NarrViz repräsentieren die Knoten die für die Geburtsberichte signifikanten n-Gramme. Sie sind in Spalten von links nach rechts angeordnet, jede Spalte (mit Farbcodierung) entspricht ein relativer Positionsbereich in den Geschichten. Knoten links stehen also am Anfang der Geschichten, jene rechts am Ende. Die Größe der Knoten steht für die Frequenz des n-Gramms. Signifikante Assoziationen zwischen den n-Grammen sind durch die Kanten dargestellt.

Die Visualisierung ist interaktiv: Bei Berührung eines Knotens werden verschiedene Informationen zum n-Gramm angezeigt. Abbildung 2.12 zeigt den ausgewählten Knoten zum n-Gramm (in lemmatisierter Form) *ich haben die Gefühl*, mit der Angabe der häufigsten Exemplare des abstrakten n-Gramms (*Ich hatte das Gefühl*, als häufigste Form). Ebenso angegeben sind statistische

**14** Die Anwendung kann unter https://pub.cl.uzh.ch/projects/visuallinguistics/NarrViz/ (letzter Zugriff: 6. 11. 2017) ausprobiert werden.


**Abb. 2.11:** Visual-Analytics-Tool „NarrViz“, Oberfläche mit Datensatz Geburtsberichte.

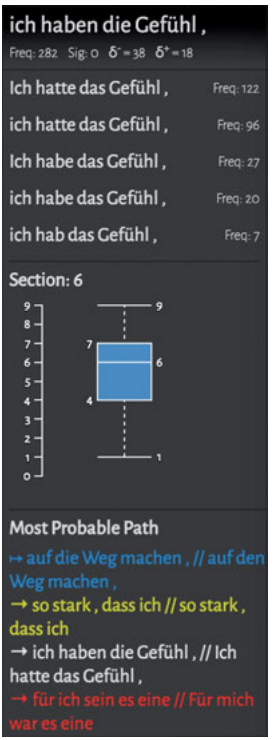


**Abb. 2.12:** Visual-Analytics-Tool „NarrViz“, Informationen zu einem n-Gramm.

Maße zum Knoten: Frequenz, Signifikanzniveau des n-Gramms (im Vergleich zum Referenzkorpus) und die Gradzentralität, also die Anzahl eintreffender (in-degree, $\delta^-$) und ausgehender (out-degree, $\delta^+$) Verbindungen. Darunter ist die Verteilung des n-Gramms über die Positionen in den Geschichten als Boxplot dargestellt: Bei der Annahme von zehn Teilen befindet sich das n-Gramm im Mittel im sechsten Teil, wobei die Hälfte der Menge zwischen dem vierten und siebten Teil streut. Ausreißer gibt es aber in allen Teilen, außer dem ersten.

Unterhalb des Boxplots zur Verteilung ist der wahrscheinlichste Pfad, also die wahrscheinlichste Sequenz aufgeführt, in der das n-Gramm vorkommt: *auf den Weg machen*, → *so stark, dass ich* → *Ich hatte das Gefühl*, → *Für mich war es eine*.[15]

Ist einer der Knoten ausgewählt, wird dieser Knoten zusammen mit seinen Verbindungen zu den anderen Knoten in der Netzwerkvisualisierung entsprechend hervorgehoben. Daneben gibt es umfangreiche Möglichkeiten, die Daten zu filtern und die Darstellung zu beeinflussen. Dies ist notwendig, da die sich ergebenden Graphen je nach Datengrundlage unterschiedlich komplex werden und bei sehr komplexen Daten gefiltert werden muss, um vernünftig arbeiten zu können.[16]

Nicht alle n-Gramme sind bezüglich ihrer Position stabil. Die Darstellung als Knoten an einer bestimmten Position ist für solche n-Gramme deshalb irreführend. Daher gibt es in NarrViz die Möglichkeit, Knoten ab einem festlegbaren Schwellwert der Standardabweichung von der mittleren Position als Balken darzustellen, deren Länge das obere und untere Quartil des Streubereichs und deren Höhe die Frequenz darstellt (vgl. Abb. 2.13). Damit ergibt sich im Überblick das Bild von n-Grammen, die sozusagen die Basis der Geschichten darstellen und an verschiedenen Positionen auftauchen. Davon heben sich dann die positionsspezifischen n-Gramme ab.

Für die linguistische Interpretation bietet es sich an, linguistische Erzähltheorien heranzuziehen, etwa jene von Labov & Waletzky (1973). Die gefundenen n-Gramme lassen sich relativ gut den dort vorgeschlagenen Erzählfunktionen zuordnen:

– **Orientierung**: *der (eigentliche) ET war der*; *gegen 22:30 Uhr sind wir*; *rief ich meinen Mann an* etc.

---

**15** Der besseren Lesbarkeit wegen werden neben der lemmatisierten, abstrakteren Form (für die die Angaben gelten) auch noch jeweils die häufigsten Wortform-Varianten angegeben.

**16** Dies widerspricht natürlich dem Paradigma der Visual Analytics, auch komplexe Daten visualisieren zu können. Die Visualisierungslösung ist deswegen auch nicht perfekt und taugt nur für Daten bis zu einem gewissen Komplexitätsgrad. Dazu kommen auch Performance-Probleme, die durch Optimierung der Implementierung behoben werden müssten.

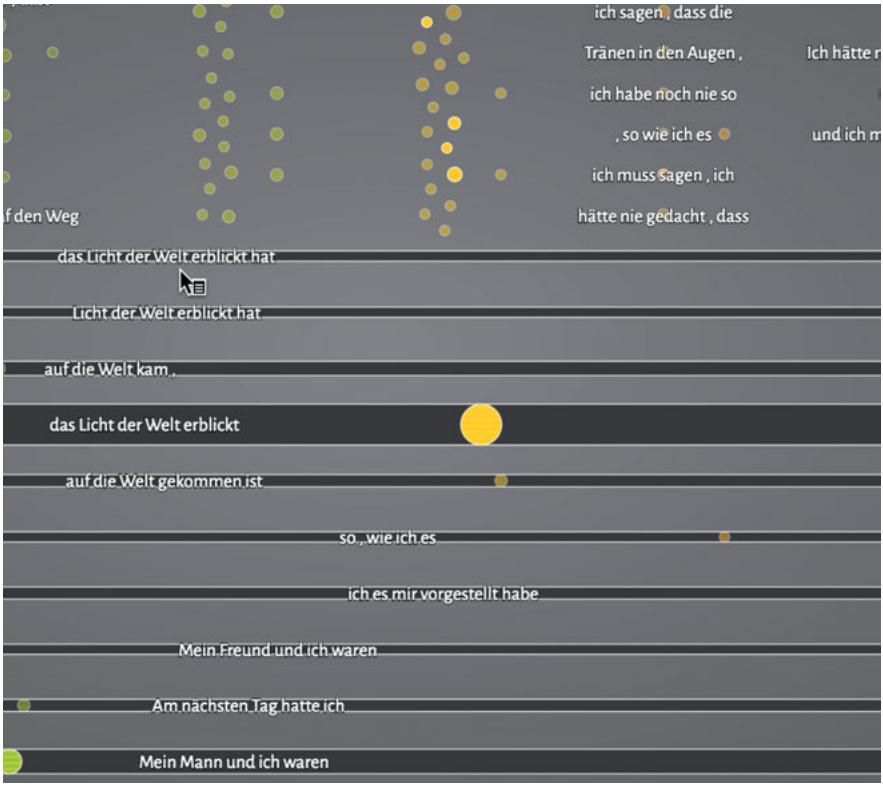


**Abb. 2.13:** Visual-Analytics-Tool „NarrViz“, Darstellung von stabilen (Punkte, oben) und variablen (Balken mit Punkten, unten) n-Grammen.

- **Komplikation**: *wurde ich ans CTG angeschlossen*; *nach langem hin und her*; *eine Wehe nach der anderen*; *wurden die Wehen immer stärker*; *Dann kam die Ärztin und* etc.
- **Evaluation**: *ich war fix und fertig*; *fühlte sich an, als*; *ich fing an zu weinen*; *ich zitterte am ganzen Körper*; *ich hatte in der ganzen Zeit [das Gefühl]* etc.
- **Resolution**: *und dann ging es los*; *ging alles sehr schnell*; *Um 15.00 Uhr war dann*; *um 16:02 Uhr das Licht*; *auf den Bauch gelegt*; *mir auf die Brust gelegt* etc.
- **Coda**: *ich bin so froh, dass*; *alles in allem war es*; *Geburt noch vor sich haben*; *der schönste Moment in meinem* etc.

Gleichzeitig lassen sich aus den Daten aber diese Kategorien auch anreichern mit auffallenden Topoi und Erzählfiguren. So ist für die Geschichten ein Moment der „Akzeleration“ typisch, das als Beginn der Resolution angesehen kann: Am

Punkt der größten Krise kurz vor der eigentlichen Austreibung ist meist ein Moment narrativer Reflexion beobachtbar (typisch für die „Evaluation" nach Labov & Waletzky 1973), bei dem das Umfeld verschwindet und das Ich der Erzählerin und ihre Gefühle im Vordergrund stehen. Dieser Stillstand wird dann aber mit n-Grammen wie *dann ging alles ganz schnell*; *dann ging alles sehr schnell* in der Erzählung aufgehoben und die Narration akzeleriert. Damit ist narrativ die Geburt mehr oder weniger bewältigt, was in der Realität natürlich nicht zwingend so ist. Die Nachgeburt oder die Wundversorgung im Anschluss sind in den Berichten aber folgerichtig auch oft kein (großes) Thema mehr. Zusammen mit der Beobachtung, dass die Geschichten meist auch gleich nach der Austreibung des Kindes mit n-Grammen wie *um 15.00 Uhr war dann*; *wurde am 29.5.2010 um 02.23 Uhr*; *um 2.24 uhr auf die welt* sozusagen die formelle Geburtsanzeige simulieren, macht die Divergenz zwischen Erlebnis und Narration deutlich: Der genaue Zeitpunkt der Geburt spielte bei der tatsächlichen Geburt wahrscheinlich für die Mutter keine Rolle, wird hinterher jedoch als wichtiger Dreh- und Angelpunkt der Geschichte konstruiert.

## 2.3 Transformationen und Operationen

Ich kann an dieser Stelle nicht weiter auf die inhaltliche Analyse eingehen[17] und möchte stattdessen einige diagrammatische Überlegungen anstellen. Bei beiden Fallbeispielen (Geokollokationen und Narrative) sind die Visualisierungen eine wichtige Hilfe für die Datenexploration. Sie hängen mit grundsätzlichen Transformationen zusammen, mit denen der Untersuchungsgegenstand erzeugt wurde. Mir scheinen insbesondere vier Grundtransformationen relevant zu sein: Rekontextualisierung, Desequenzialisierung, Dimensionsanreicherung und Rematerialisierung. Was ist damit gemeint?

Jeglicher quantitativer Korpusanalyse eigen ist eine *Rekontextualisierung* von sprachlichen Einheiten: Eine KWiC-Liste ist ein Ensemble von Einzelbelegen, die aus ihren ursprünglichen Kontexten extrahiert und zu einer Liste rekontextualisiert worden sind, die den Untersuchungsgegenstand darstellt. Die Fundstellen werden nicht mehr als Funde innerhalb eines Textes gelesen, sondern als Ensemble aller Fundstellen. Die Einheit des Textes wurde zerstört, um eine Perspektive zu ermöglichen, die nach der Musterhaftigkeit der Verwendung dieser Funde fragt. Bei den Geburtsberichten erzeugten wir durch die datengeleitete Berechnung von typischen n-Grammen den Untersuchungsgegenstand, mit dem die n-Gramme als musterhafte n-Gramme rekontextuali-

**17** Vgl. dazu Bubenhofer (2018a).

siert werden: Sie sind ihren ursprünglichen Kontexten entrissen, zeigen dafür die Musterhaftigkeit ihrer Verwendung – erstens weil es sich um Wortsequenzen handelt, die in den Daten gehäuft in diesen Gruppen auftreten, und zweitens weil sie bezüglich ihrer Frequenz auffällig oft in den Geburtsberichten vorkommen. Bei den Geokollokationen hingegen ist das Kollokationsprofil zu jedem Toponym eine Rekontextualisierung von Lexemen im Umfeld des jeweiligen Toponyms, wobei das Kollokationsprofil die Verwendungsweisen kompakt zusammenfasst.

Mit der Rekontextualisierung gehen *Desequenzialisierungen* einher: Bei den Kollokationsprofilen der Geokollokationen sind überhaupt keine Informationen über die syntagmatische Einbettung in den Kontext verfügbar. Die Kombination aus Rekontextualisierung und Desequenzialisierung ist unter korpuslinguistisch-diskursanalytischer Perspektive ein erwünschter Effekt:

> Die dekontextualisierte Darstellung erlaubt es den Forschenden, frei vom ‚hermeneutischen Reflex', der die Lektüre von Texten und Textpassagen bestimmt, kreativ Ideen zu möglichen diskursiven Zusammenhängen einzelner Korpusteile zu entwickeln, die bei einer subjektiven Lektüre möglicherweise verdeckt blieben. (Scholz & Mattissek 2014: 87)

Es ist also einerseits gerade notwendig, vom Einzeltext zu extrahieren (was das Ziel aller quantitativen Analysen ist), andererseits aber auch von der sequenziellen Einbettung der Belege.[18]

Bei den Narrativen ist die Desequenzialisierung kritisch, weil die Abfolge der n-Gramme im Verlauf der Geschichte von großem Interesse ist. Allerdings interessiert uns nicht die einzelne Geschichte, sondern die generalisierte, der musterhafte Ablauf. Daher ist es wichtig, die entsprechenden Daten der typischen Sequenzen zu erheben und die Visualisierung an der Grundfigur der Sequenz auszurichten.

Die dritte relevante Grundfigur ist die *Dimensionsanreicherung*: Bei den Geokollokationen wird durch die Georeferenzierung eine weitere Dimension hinzugefügt mit dem Ziel, eine Visualisierung zu ermöglichen, die die Daten unter einer neuen Perspektive interpretierbar macht. Ich habe auch problematisiert, dass die Kartendarstellung kritisch ist, da es sich um eine kanonische Form der Dimensionsanreicherung handelt, die oft vorschnell als einzig relevante gewählt wird. Daher ist es wichtig, mit unterschiedlichen Anreicherungen zu arbeiten, was wir in Form des Dorling-Diagramms versucht haben.

---

**18** Ein Rest an (musterhafter) syntagmatischer Einbettung ist durch die Berechnung der Kollokation natürlich noch vorhanden. Vgl. für eine weiterführende Diskussion auch Bubenhofer (2018c, b).

Bei den Narrativen stellt die Berechnung der typischen Positionen der n-Gramme in den Geschichten die wichtige Dimensionsanreicherung dar, mit der es möglich wird, die typischen narrativen Sequenzen zu finden. Sieht man diese Positionierung der n-Gramme aber auch nur als eine von vielen möglichen Anreicherungen, wird klar, dass auch nach Alternativen gesucht werden muss. Schließlich erzeugen die diagrammatischen Transformationen eine Form von *Rematerialisierung*. Darunter verstehe ich die eigentliche Konstitution des Untersuchungsgegenstands auf einer emergenten Ebene: Ein Kollokationsprofil ist beispielsweise eine statistische Zusammenfassung des Distributionsverhaltens eines Ausdrucks, das durch diagrammatische Transformationen entstanden ist. Bei der Analyse behandeln wir das Profil als Analysematerial, das semantische Lesarten des Lexems darstellt. Ähnlich die sprachlichen Einheiten, die georeferenziert und auf einer Karte dargestellt werden: Sie ergeben einen Gegenstand von Sprache „in situ“. Wir können die diagrammatisch manipulierten Daten auf einer emergenten Ebene als einen neuen Gegenstand lesen und interpretieren.

Das Beispiel der Geokollokationen zeigt eine Rematerialisierung als diskursives Bedeutungsgewebe zur Konstruktion von Welt. Dieses Bedeutungsgewebe bewegt sich dabei zwischen stark und weniger stark geografisch verankerten Formen: Bei der Rückbindung auf die Kartenprojektion sind die Abweichungen zwischen diskursivem Weltbild und geografischem besonders deutlich sichtbar. Die Dorling-Visualisierung hingegen spiegelt eher das diskursive Weltbild.

Bei den Narrativen wird durch die Visualisierung die Musterhaftigkeit der typischen Verkettungen gezeigt, also eine Abstrahierung auf zwei Ebenen: Einerseits sind die statistisch auffälligen Sequenzen sichtbar, andererseits aber wiederum eine Abstrahierung der einzelnen auffälligen Sequenzen auf wenige narrative Muster. Diese narrativen Muster erzählen eine Geschichte, wie sie genau so nicht in den Daten zu finden ist, jedoch trotzdem eine passende Typisierung darstellt. Sie lautet z. B. so:

> An diesem Tag hatte ich … → …, dass es endlich losgeht → ich hatte das Gefühl, dass … → Mein Mann und ich waren … → war mich sicher, dass … → auf den Weg in die … → Ich sagte ihr, dass … → so heftig, dass ich … → fühlte sich an, als → war ich fix und fertig → Ich hatte das Gefühl, → dass es nicht mehr lange … → ich dachte, ich muss … → , aber es ging nicht → , was das Zeug hielt → dann ging alles ganz schnell → Ich weiß nur noch … → um 16:38 war es → das Licht der Welt erblickte → ich konnte es nicht glauben → ich war so froh … → , dass es vorbei war → ich hätte nie gedacht, → und ich muss sagen, → Für mich war es eine … → noch vor sich haben …

# 3 Plädoyer für mehr Experimentierfreude

Auch die Sozial- und Kulturwissenschaften tendieren dazu, Standardmethoden zu entwickeln, um bestimmte Analysen valide und reliabel durchführen zu können. Das führt z. B. zu Best Practice-Empfehlungen für die Datenanalyse und Datenvisualisierung, die Eingang in korpuslinguistische Literatur finden. Dies ist wichtig. Trotzdem: Die grundlegenden Überlegungen zu diagrammatischen Operationen und Grundfiguren, die etwa die Korpuslinguistik beherrschen, machen deutlich, wie grundlegend bestimmte methodische Zugriffe bei der Gegenstandskonstitution sind. Methoden der Visualisierung, egal ob für explorative Zwecke oder für die Präsentation, sind nicht einfach ein zusätzliches Element der Analyse, sondern prägen ganz grundsätzlich unsere Perspektive auf die Daten. Dies wurde in wissenschaftstheoretischen Arbeiten, insbesondere zu den Naturwissenschaften, verschiedentlich aufgezeigt (Knorr Cetina 2001; Böhm 2001; Rheinberger 1994; Bredekamp, Schneider & Dünkel 2008; Zittel 2011; Burri 2016). In der Linguistik sind die Reflexionen darüber, wie die im Fach herrschenden Diagrammkulturen die Gegenstandskonstitution „Sprache" prägen, noch wenig erforscht. Gerade die Korpuslinguistik muss mit ihrem datengeleiteten Zugriff eine Vorreiterrolle übernehmen, um nicht voreilig Standards der Visualisierung zu formulieren, sondern mit verschiedenen Formen zu experimentieren und diese auch wissenschaftstheoretisch zu reflektieren. Mit Lauersdorf (2018) kann man fordern: „Use all the data! View all the data! View all the combinations! View all the angles! Use all the techniques!"

Dies kann beispielsweise auch in Domänen geschehen, die auf den ersten Blick vielleicht wenig Berührungspunkte zur Korpuslinguistik haben. Ein Beispiel ist die Gesprächslinguistik oder interaktionale Linguistik. Das Gesprächstranskript spielt hier die entscheidende Rolle, um gesprochene Sprache überhaupt zu Daten zu machen, die analysiert werden können (z. B. nach GAT; Selting et al. 1998). Gespräche können jedoch auch ganz anders visualisiert werden, wie der Versuch zeigt, Gespräche mit der Figur der Jahresringe zu visualisieren (vgl. Abb. 2.14). Die grafische Figur sieht vor, die Aktanten eines Gesprächs als Positionen auf einem Kreis zu sehen. Parallel zum Ablauf des Gesprächs (die Aufnahme wird abgespielt), produziert jeder Turn an der Stelle des Aktanten einen Kreis, wobei die Länge des Turns die Größe bestimmt. Konzentrisch dazu werden die weiteren Turns des Aktanten gezeichnet, wobei die Kreise unmittelbar bei turn-Äußerung farbig gefüllt sind, mit der Zeit jedoch verblassen, aber nicht unsichtbar werden. Zusätzlich wird der Text des jeweiligen Turns kurz angezeigt.

Nach einiger Zeit wird eine Geschichte und Dynamik des Gesprächs sichtbar, wobei Abbildung 2.14 die Unterschiede dreier Gespräche augenfällig

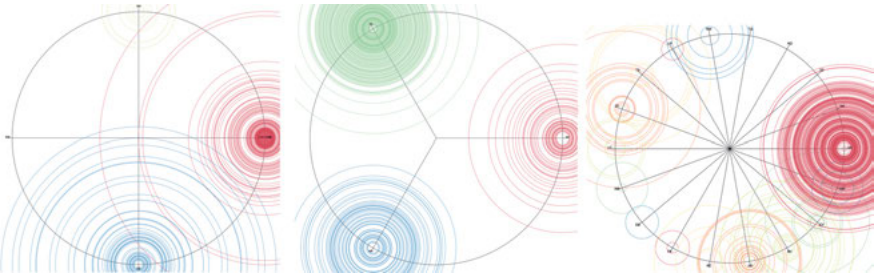

**Abb. 2.14:** Visualisierung der Gesprächsdynamik mit „Jahresringen“ – Gespräche 1–3.

macht: Das erste Gespräch (in Abb. 14 links) vollzieht sich hauptsächlich zwischen zwei Personen, die beide viele, relativ kurze Turns geäußert haben, daneben aber auch längere in ähnlicher Zahl. Man würde sagen, es handelt sich um eine ausgeglichene Dialogform. Beim zweiten Gespräch (in Abb. 14 in der Mitte) nehmen drei Personen am Gespräch teil, wobei der eine (links oben) das Gespräch mit sehr vielen, eher kurzen turns dominiert. In der dritten Konstellation (in Abb. 14 rechts) agieren viele Personen miteinander, wobei die eine ebenfalls das Gespräch deutlich dominiert, gefolgt von drei, vier weiteren Personen, die auch substanziell zum Gespräch beitragen.

Neue Formen der Datenvisualisierung sind nötig, da die theoretischen Überlegungen zur interaktionalen Linguistik weiter zu sein scheinen als die gängigen Formen der Gesprächstranskription. So sieht Deppermann (2014: 323) „vier Bestimmungsstücke des sprachlichen Handelns [die] unser Verständnis von ‚Pragmatik‘ prägen müssen: Leiblichkeit [...], Zeitlichkeit [...], Sozialität [...], Epistemizität“. Zeitlichkeit meint dabei, dass sprachliches Handeln „sequenziell organisiert und simultan mit anderen Ressourcen des Handelns verknüpft“ ist und deshalb „Retrospektion und Projektion [...] konstitutive Dimensionen der situierten Sinnkonstitution“ sind. Anhand eines klassischen Gesprächstranskripts ist es sehr schwer, Retrospektion und Projektion abzuleiten. Der „Jahresringe“-Darstellung in Abbildung 2.14 ist aber immerhin ein Element der Retrospektion eingeschrieben: Eine Geschichte des Gesprächs ergibt sich in grafischer Form. Auch der Aspekt der Sozialität – „[s]prachliches Handeln findet in interpersonellen (Mehrpersonen-)Konstellationen statt“ (Deppermann 2014: 323) – ist in der Darstellung sichtbar.

Die „Jahresringe“-Darstellung ist nur eine erste Skizze für eine die klassischen Transkriptionsformen ergänzende Datenvisualisierung, die stark ausgebaut werden müsste, um brauchbar zu sein. Ich wollte damit aber zeigen, dass neue Wege der Datenanalyse theoriegeleitet vorgehen und dabei auch neue diagrammatische Grundfiguren finden müssen, um neue Perspektiven

auf die Daten zu ermöglichen. Dabei ist auch klar, dass eine neue Visualisierung nicht den Zweck hat, die alten Fragen besser zu beantworten, sondern neue Fragen überhaupt erst ermöglicht. Der hier skizzierte „Jahresringe"-Vorschlag ist dabei zugegebenermaßen von einem deutlich korpuslinguistischen Blick auf gesprochene Sprache geprägt.

Neben methodischen Standards der Datenaufbereitung und Analyse ist es aber gerade auch in der Korpuslinguistik ein Desiderat, Experimente der Datenvisualisierung einzugehen und dabei von einer weitreichenden Definition von Diagramm auszugehen. Die Visualisierung beginnt nicht erst mit der Analyse (oder gar der Präsentation der Analyseergebnisse). Diagrammatische Überlegungen gehen bereits mit dem theoretischen Zugriff auf die Daten einher und sind ein wichtiges Element der Gegenstandskonstitution.

## Literatur


Affolter, Katrin (2016): *Visualization of narrative structures*. Universität Zürich Master-Arbeit.

Barbieri, Gian Luca, Ada Cigala, Alessandro Musetti & Paolo Corsano (2012): Looking forward to the birth of a child: Tales of motherhood in forums. *International Journal of Psychoanalysis and Education IJPE* 4(2), 4–26.

Bauer, Matthias & Christoph Ernst (2010): *Diagrammatik / Einführung in ein kultur- und medienwissenschaftliches Forschungsfeld*. Bielefeld: transcript.

Bański, Piotr (2010): Why TEI stand-off annotation doesn't quite work: And why you might want to use it nevertheless. In *Balisage: The Markup Conference 2010*, Band 5, Montréal, Canada. doi: 10.4242/BalisageVol5.Banski01.

Belica, Cyril (2001 ff.): *Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemischstrukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs*. Mannheim: Institut für Deutsche Sprache. http://corpora.idsmannheim.de (letzter Zugriff: 6. 11. 2017).

Bender, John B. & Michael Marrinan (2014): *Kultur des Diagramms* (Actus et imago Band VIII). Berlin: Akademie Verlag.

Bonfanti, Corrado (2012): Roberto Busa (1913–2011), pioneer of computers for the humanities. In Arthur Tatnall (Hrsg.), *Reflections on the history of computing. Preserving memories and sharing stories*, 57–61. Heidelberg: Springer.

Bostock, Michael, Vadim Ogievetsky & Jeffrey Heer (2011): D3: data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* http://vis.stanford.edu/ papers/d3 (letzter Zugriff: 6. 11. 2017).

Bredekamp, Horst (2008): Diagrammatik. In Horst Bredekamp, Birgit Schneider & Vera Dünkel (Hrsg.), *Das Technische Bild: Kompendium zu einer Stilgeschichte wissenschaftlicher Bilder*, 192–197. Berlin: Akademie-Verlag.

Bredekamp, Horst, Birgit Schneider & Vera Dünkel (Hrsg.) (2008): *Das Technische Bild: Kompendium zu einer Stilgeschichte wissenschaftlicher Bilder*. Berlin: Akademie Verlag.

Brezina, Vaclav, Tony McEnery & Stephen Wattam (2015): Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2), 139–173. doi: 10.1075/ijcl.20.2.01bre.

Bubenhofer, Noah (2014): Geokollokationen – Diskurse zu Orten: Visuelle Korpusanalyse. *Mitteilungen des Deutschen Germanistenverbandes* 61(1), 45–59.

Bubenhofer, Noah (2017): Kollokationen, n-Gramme, Mehrworteinheiten. In Kersten Roth, Martin Wengeler & Alexander Ziem (Hrsg.), *Handbuch Sprache in Politik und Gesellschaft* (Handbücher Sprachwissen 19), 69–93. Berlin, Boston: de Gruyter Mouton.

Bubenhofer, Noah (2018a): Serialität der Singularität. Korpusanalyse narrativer Muster in Geburtsberichten. *Zeitschrift für Literaturwissenschaft und Linguistik: Themenheft Alltagspraktiken des Erzählens*.

Bubenhofer, Noah (2018b): Diskurslinguistik und Korpora: Daten im Vektorraum. In Ingo Warnke (Hrsg.), *Handbuch Diskurs* Handbücher Sprachwissen, Berlin, Boston: de Gruyter Mouton.

Bubenhofer, Noah (2018c): Visual Linguistics: Plädoyer für ein neues Forschungsfeld. In Noah Bubenhofer & Marc Kupietz (Hrsg.), Visualisierung sprachlicher Daten: Visual Linguistics –Praxis – Tools. Heidelberg: Heidelberg University Publishing. doi: 10.17885/heiup.345.474

Bubenhofer, Noah, Nicole Müller & Joachim Scharloth (2013): Narrative Muster und Diskursanalyse: Ein datengeleiteter Ansatz. *Zeitschrift für Semiotik, Methoden der Diskursanalyse* 35(3–4), 419–444.

Bubenhofer, Noah & Klaus Rothenhäusler (2016): „Korporatheken“: Die digitale und verdatete Bibliothek. *027.7 Zeitschrift für Bibliothekskultur/Journal for Library Culture* 4(2), 60–71.

Bubenhofer, Noah, Klaus Rothenhäusler, Katrin Affolter & Danica Pajovic (2017): The linguistic construction of world – an example of visual analysis and methodological challenges. In Ronny Scholz (Hrsg.), *Quantifying approaches to discourse for social scientists*, Basingstoke: Palgrave Macmillan.

Burri, Regula Valérie (2016): Bilder als soziale Praxis: Grundlegungen einer Soziologie des Visuellen/Images as social practice: Outline of a sociology of the visual. *Zeitschrift für Soziologie* 37(4), 342–358. doi: 10.1515/zfsoz-2008-0404.

Busa, Roberto (1951): *Sancti Thomae Aquinatis Hymnorum ritualium varia specimina concordantiarum: primo saggio di indici di parole automaticamente composti e stampi da macchine IBM a schede perforate = A 1st example of word index automatically compiled and printed by IBM punched card machines* (Archivum philosophicum Aloisianum. Serie 2). Milano: Bocca.

Böhm, Gottfried (2001): Zwischen Auge und Hand. Bilder als Instrumente der Erkenntnis. In Bettina Heintz & Jörg Huber (Hrsg.), *Mit dem Auge denken: Strategien der Sichtbarmachung in wissenschaftlichen und virtuellen Welten* (Theorie:Gestaltung 01), 43–54. Wien, New York: Springer.

Chen, Chun-houh, Wolfgang Härdle & Antony Unwin (Hrsg.) (2008): *Handbook of data visualization* (Springer handbooks of computational statistics). Heidelberg: Springer.

Colloseus, Cecilia (2016): *Gebären – Erzählen. Kulturanthropologische und interdisziplinäre Perspektiven auf die Geburt als leibkörperliche Grenzerfahrung*. Mainz: Johannes Gutenberg-Universität Dissertation.

Deppermann, Arnulf (2014): Pragmatik revisited. In Ludwig Eichinger (Hrsg.), *Sprachwissenschaft im Fokus Positionsbestimmungen und Perspektiven* Jahrbuch des Instituts für Deutsche Sprache, 323–352. Berlin, Boston: De Gruyter.

Dorling, Danny (1993): Map design for census mapping. *The Cartographic Journal* 30(2), 167–183. doi: 10.1179/000870493787860175.

Echterhölter, Anna (2015): Jack Goody: Die Liste als Praktik. In Susanne Deicher & Erik Maroko (Hrsg.), *Die Liste: Ordnungen von Dingen und Menschen in Ägypten*, Band 1: Ancient Egyptian design, contemporary design history and anthropology of design, 243–261. Berlin: Kulturverlag Kadmos.

Eco, Umberto (1977): *Zeichen. Einführung in einen Begriff und seine Geschichte*. Frankfurt am Main: Suhrkamp.

Eco, Umberto (2009): *Die unendliche Liste*. München: Carl Hanser.

Ehlich, Konrad (Hrsg.) (1980): *Erzählen im Alltag* (Suhrkamp-Taschenbuch Wissenschaft 323). Frankfurt am Main: Suhrkamp.

Evert, Stefan (2009): Corpora and collocations. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus linguistics*, Band 2 (Handbücher zur Sprach- und Kommunikationswissenschaft 29.2), 1212–1248. Berlin, Boston: Mouton de Gruyter.

Faruqui, Manaal & Sebastian Padó (2010): Training and evaluating a german named entity recognizer with semantic generalization. In Manfred Pinkal, Ines Rehbein, Sabine Schulte im Walde & Angelika Storrer (Hrsg.), *Proceedings of KONVENS, September 6–8, 2010, Saarland University, Saarbrücken*, 129–134. Saarbrücken.

Finkel, Jenny Rose, Trond Grenager & Christopher Manning (2005): Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the ACL*, 363–370. Stroudsburg, PA: Association for Computational Linguistics.

Glasze, Georg (2009): Kritische Kartographie. *Geographische Zeitschrift* 97(4), 181–191.

Graham, Shawn, Scott Weingart & Ian Milligan (2012): Getting started with topic modeling and MALLET. http://programminghistorian.org/lessons/topic-modeling-and-mallet (letzter Zugriff: 6. 11. 2017).

Gülich, Elisabeth (1980): Konventionelle Muster und kommunikative Funktionen von Alltagserzählungen. In Konrad Ehlich (Hrsg.), *Erzählen im Alltag* (Suhrkamp-Taschenbuch Wissenschaft 323), 335–384. Frankfurt am Main: Suhrkamp.

Jullien, François (2004): Die praktische Wirkkraft der Liste: von der Hand, vom Körper, vom Gedicht. In François Jullien (Hrsg.), *Die Kunst, Listen zu erstellen*, 15–50. Berlin: Merve.

Keim, Daniel A., Jörn Kohlhammer, Geoffrey Ellis & Florian Mansmann (2010): *Mastering the information age – solving problems with visual analytics*. Goslar: Eurographics Association.

Kleiner, Stefan (2011 ff.): *Atlas zur Aussprache des deutschen Gebrauchsstandards (AADG)*. Mannheim: IDS. http://prowiki.ids-mannheim.de/bin/view/AADG/ (letzter Zugriff: 6. 11. 2017).

Knorr Cetina, Karin (2001): „Viskurse" der Physik. Konsensbildung und visuelle Darstellung. In Bettina Heintz & Jörg Huber (Hrsg.), *Mit dem Auge denken: Strategien der Sichtbarmachung in wissenschaftlichen und virtuellen Wellen* (Theorie:Gestaltung 01), 305–320. Wien, New York: Springer.

Koplenig, Alexander (2017): A data-driven method to identify (correlated) changes in chronological corpora. *Journal of Quantitative Linguistics* 4(24), 289–318. doi: 10.1080/09296174.2017.1311447.

Kruja, Eriola, Joe Marks, Ann Blair & Richard Waters (2002): A short note on the history of graph drawing. In Petra Mutzel, Michael Jünger & Sebastian Leipert (Hrsg.), *Graph drawing* (Lecture Notes in Computer Science 2265), 272–286. Berlin, Heidelberg: Springer.

Krämer, Sybille (2016): *Figuration, Anschauung, Erkenntnis: Grundlinien einer Diagrammatologie*. Frankfurt am Main: Suhrkamp.

Labov, William & Joshua Waletzky (1973): Erzählanalyse. Mündliche Versionen persönlicher Erfahrung. In Jens Ihwe (Hrsg.), *Literaturwissenschaft und Linguistik*, Band 2, 78–126. Frankfurt am Main: Athenäum.

Lauersdorf, Mark Richard (2018): Linguistic visualizations as objets d'art? In Noah Bubenhofer & Marc Kupietz (Hrsg.), *Visual linguistics*, Heidelberg: Heidelberg University Publishing. [Im Druck].

Lima, Manuel (2014): *The book of trees: Visualizing branches of knowledge*. New York: Princeton Architectural Press.

Luhn, Hans Peter (1960): Key word-in-context index for technical literature (kwic index). *American Documentation* 11(4), 288–295. doi: 10.1002/asi.5090110403.

Manning, Christopher D. & Hinrich Schütze (2002): *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean (2013): Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs]* http://arxiv.org/abs/1301.3781 (letzter Zugriff: 6. 11. 2017).

Naumann, Carl Ludwig (1982): Kartographische Datendarstellung. In Werner Besch, Ulrich Knoop, Wolfgang Putschke & Herbert E. Wiegand (Hrsg.), *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, Band 1 (Handbücher zur Sprach- und Kommunikationswissenschaft 1), 667–692. Berlin, Boston: de Gruyter Mouton.

Peirce, Charles S. (1994): *The collected papers of Charles Sanders Peirce*. Charlottesville, VA: InteLex Corp. http://www.nlx.com/collections/95 (letzter Zugriff: 6. 11. 2017).

Perkuhn, Rainer & Cyril Belica (2006): Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik. *Sprachreport* 22(1), 2–8.

Perkuhn, Rainer, Holger Keibel & Marc Kupietz (2012): *Korpuslinguistik*. Stuttgart: UTB.

Pfeffer, Jürgen (2010): Visualisierung sozialer Netzwerke. In Christian Stegbauer (Hrsg.), *Netzwerkanalyse und Netzwerktheorie*, 227–238. Wiesbaden: Springer VS.

Pigeot, Jacqueline (2004): Die explodierte Liste: die Tradition der heterogenen Liste in der alten japanischen Literatur. In François Jullien (Hrsg.), *Die Kunst, Listen zu erstellen*, 73–121. Berlin: Merve.

Placcius, Vincentius (1689): *De arte excerpendi. Vom Gelahrten Buchhalten.* Hamburg: Gottfried Liebezeit. http://echo.mpiwg-berlin.mpg.de/MPIWG:7X92X29A (letzter Zugriff: 6. 11. 2017).

Quasthoff, Uta (1980): *Erzählen in Gesprächen: linguistische Untersuchungen zu Strukturen und Funktionen am Beispiel einer Kommunikationsform des Alltags* (Kommunikation und Institution 1). Tübingen: Narr.

Redder, Angelika (2001): Aufbau und Gestaltung von Transkriptionssystemen. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann & Sven Sager (Hrsg.), *Text- und Gesprächslinguistik / Linguistics of text and conversation*, Band 2 (Handbücher zur Sprach- und Kommunikationswissenschaft 16.2), 1038–1059. Berlin, Boston: de Gruyter Mouton.

Rheinberger, Hans-Jörg (1994): Experimentalsysteme, Epistemische Dinge, Experimentalkulturen Zu einer Epistemologie des Experiments. *Deutsche Zeitschrift für Philosophie* 42(3), 405–418.

Sachs, Klaus-Jürgen & Thomas Röder (1989): Partitur. In Ludwig Finscher (Hrsg.), *Die Musik in Geschichte und Gegenwart: allgemeine Enzyklopädie der Musik*, Band 10, 1424–1437.

Sacks, Harvey, Emanuel A. Schegloff & Gail Jefferson (1974): A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4), 696–735. doi: 10.2307/412243.

Sankey, Henry Riall (1896): The thermal efficiency of steam-engines. (including appendixes). *Minutes of the Proceedings of the Institution of Civil Engineers* 125(1896), 182–212. doi: 10.1680/imotp.1896.19564.

Schiller, Anne, Simone Teufel & Christine Thielen (1995): *Guidelines für das Tagging deutscher Textcorpora mit STTS. Arbeitspapier*. Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, Universität Tübingen, Seminar für Sprachwissenschaft.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf (letzter Zugriff: 6. 11. 2017).

Schmid, Helmut (1995): *Improvements in part-of-speech tagging with an application to german*. Dublin. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf (letzter Zugriff: 6. 11. 2017).

Scholz, Ronny & Annika Mattissek (2014): Zwischen Exzellenz und Bildungsstreik. Lexikometrie als Methodik zur Ermittlung semantischer Makrostrukturen des Hochschulreformdiskurses. In Johannes Angermüller, Martin Nonhoff, Eva Herschinger, Felicitas Macgilchrist, Martin Reisigl, Juliette Wedl, Daniel Wrana & Alexander Ziem (Hrsg.), *Diskursforschung. Ein interdisziplinäres Handbuch*, Band 2, 86–112. Bielefeld: transcript.

Schumann, Heidrun & Wolfgang Müller (1999): *Visualisierung: Grundlagen und allgemeine Methoden*. Springer DE.

Selting, Margret, Peter Auer, Birgit Barden, Jörg Bergmann, Elizabeth Couper-Kuhlen, Susanne Günthner, Christoph Meier, Uta Quasthoff, Peter Schlobinski & Susanne Uhmann (1998): Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte* 173, 91–122.

Siegel, Steffen (2009): *Tabula: Figuren der Ordnung um 1600*. Berlin: Akademie Verlag.

Smith, Neil (1992): History and philosophy of geography: Real wars, theory wars. *Progress in Human Geography* 16(2), 257–271. doi: 10.1177/030913259201600208.

Stetter, Christian (2005): Bild, Diagramm, Schrift. In Gernot Grube, Werner Kogge & Sybille Krämer (Hrsg.), *Schrift. Kulturtechnik zwischen Auge, Hand und Maschine* (Kulturtechnik 4), 115–136. München: Wilhelm Fink Verlag.

Steyer, Kathrin (2013): *Usuelle Wortverbindungen: Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht* (Studien zur Deutschen Sprache 65). Tübingen: Narr.

Zittel, Claus (2011): Ludwik Fleck und der Stilbegriff in den Naturwissenschaften. Stil als wissenschaftshistorische, epistemologische und ästhetische Kategorie. In Horst Bredekamp & John Michael Krois (Hrsg.), *Sehen und Handeln*, 171–206. Berlin, Boston: de Gruyter.